

HYBRIDIZATION NETWORKS

**Charles Semple**

*Department of Mathematics and Statistics  
University of Canterbury  
Private Bag 4800  
Christchurch, New Zealand*

**Report Number:** UCDMS2006/3

**MAY 2006**

# HYBRIDIZATION NETWORKS

CHARLES SEMPLE

**ABSTRACT.** Reticulate evolution is a fundamental process in the evolution of certain groups of taxa. Consequently, conflicting signals in a data set may not be the result of sampling or modelling errors, but due to the fact that reticulation has played a role in the evolutionary history of the species under consideration. However, despite its occurrence, such processes are still relatively rare, and so, assuming our initial data set is correct, a fundamental problem is to compute the minimum number of reticulation events that explains this set. In this chapter, we focus our attention on this problem for when the initial set consists of two rooted binary phylogenetic trees. This may seem rather special, but there are several reasons for this. Firstly, the problem is NP-hard even when the initial set consists of two such trees. Secondly, we are interested in finding a general solution rather than one that is restricted in some way. Lastly, the problem for when the initial set consists of binary sequences can be interpreted as a sequence of two-tree problems.

## 1. INTRODUCTION

Evolutionary (phylogenetic) trees are used to represent the tree-like evolution of a collection of taxa. For many groups of taxa (for example, mammals) this representation is appropriate. However, non-tree-like evolutionary processes such as hybridization, horizontal gene transfer, and recombination mean that not all groups of taxa are suited to this type of representation. Collectively referred to as reticulation events, these types of processes result in species being a composite of genes derived from different ancestors. Such groups include certain plant and fish species.

The effect of reticulation in evolution has been recognized for quite some time. Since the 1930's, botanists suggested that the morphological variation in the New Zealand flora is due to hybridization [2]. More recently, in the context of horizontal gene transfer, Doolittle [13] wrote that "molecular phylogeneticists will have failed to find the 'true tree', not because their methods are inadequate or because they have chosen the wrong genes, but because the history of life cannot be properly represented as a tree." Despite this recognition, mathematical investigations into the understanding and analysis of reticulation in evolution are relatively recent.

---

*Date:* 3 May 2006.

1991 *Mathematics Subject Classification.* 05C05; 92D15.

*Key words and phrases.* Hybridization networks, recombination networks, rooted subtree prune and regraft, agreement forests.

This work was supported by the New Zealand Marsden Fund (UOC310).

In a separate chapter, Huson provides an overview of various ways of representing the evolutionary history of a collection of taxa that has undergone reticulate evolution. In this chapter, we focus our attention on a particular problem that is both biologically important and mathematically challenging. It is commonly accepted that reticulation is relatively rare and so a fundamental problem for biologists studying the evolution of species whose past has included reticulation is the following: given a collection of rooted phylogenetic trees on sets of species that correctly represents the tree-like evolution of different parts of their genomes, what is the smallest number of reticulation events needed to explain the evolution of the species under consideration. This smallest number sets a lower bound on the number of such events.

The chapter is organized as follows. In Section 2, we formalize the above problem and the notion of a hybridization network, the latter is central to this problem. In general, the problem is NP-hard even when the initial collection consists of two trees. However, there is an attractive and particularly useful characterization of it in this case. This characterization is described in Section 3, while Section 4 contains algorithmic applications of it. In Section 5, we consider the variant of the problem for when the initial collection is a set of binary sequences. The material in this section is used in the subsequent two sections. An important biological consideration of the evolutionary history of taxa is that reticulation events occur between taxa that coexist in time. We investigate this consideration in Section 6, while, in Section 7, we consider some of the computational issues in computing the above smallest number.

For completeness, we end this section with some preliminaries. Unless otherwise stated, the notation and terminology in this chapter follows Semple and Steel [37].

**Preliminaries.** A *rooted phylogenetic  $X$ -tree*  $\mathcal{T}$  is a rooted tree whose root has degree at least two and whose leaf set is  $X$ . In addition,  $\mathcal{T}$  is *binary* if, apart from the root which has degree two, all interior vertices have degree three. The set  $X$  is called the *label set* of  $\mathcal{T}$  and we sometimes denote it as  $\mathcal{L}(\mathcal{T})$ . Examples of rooted binary phylogenetic trees are shown in Fig. 1 and at the top of Fig. 2.

For convenience, many of the examples that arise in this chapter are based on rooted caterpillar trees. A *rooted caterpillar tree* is a rooted binary phylogenetic tree that has a leaf vertex,  $x$  say, such that every other leaf vertex is attached to the path from  $x$  to the root via a pendant edge. The rooted binary phylogenetic tree shown in Fig. 1 is an example of a rooted caterpillar tree. Without ambiguity, we denote this rooted caterpillar tree by the  $n$ -tuple  $(x_1, x_2, \dots, x_n)$  as this is the ordering of the label set induced by the path from  $x_1$  to the root. Note that the first two coordinates of this tuple could be interchanged to describe the same rooted caterpillar tree.

Let  $\mathcal{T}$  be a rooted phylogenetic  $X$ -tree and let  $v$  be a vertex of  $\mathcal{T}$ . The subset of elements  $X$  that are descendants of  $v$  is called a *cluster* of  $\mathcal{T}$ . We denote this cluster by  $C_{\mathcal{T}}(v)$  or simply  $C(v)$  if there is no ambiguity. We sometimes say that  $C(v)$  is the cluster of  $\mathcal{T}$  *corresponding to  $v$  in  $\mathcal{T}$* . The set of clusters of  $\mathcal{T}$  is denoted by  $\mathcal{C}(\mathcal{T})$ . Note here that the root of  $\mathcal{T}$  gives rise to a cluster. This differs to the

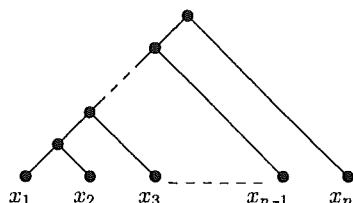


FIGURE 1. A rooted caterpillar tree.

definition in [37] where, for technical reasons, a cluster is not associated with the root of  $\mathcal{T}$ .

For a rooted phylogenetic  $X$ -tree  $\mathcal{T}$ , several different types of rooted subtrees will play a prominent role in this chapter. Let  $X'$  be a subset of  $X$ . The minimal rooted subtree of  $\mathcal{T}$  that connects the leaves in  $X'$  is denoted by  $\mathcal{T}(X')$ . Furthermore, the *restriction* of  $\mathcal{T}$  to  $X'$ , denoted by  $\mathcal{T}|X'$ , is the rooted phylogenetic tree obtained from  $\mathcal{T}(X')$  by suppressing any non-root vertices of degree 2. Lastly, a rooted subtree of  $\mathcal{T}$  is *pendant* if it can be obtained from  $\mathcal{T}$  by deleting a single edge.

## 2. HYBRIDIZATION NETWORKS

In this section, we formalize the optimization problem described in the introduction. We begin with the concept of a hybridization network which is central to this problem and this chapter. These networks are particular types of digraphs.

A *directed graph* (also known as a *digraph*) consists of a collection of vertices and a collection of directed edges called *arcs*. If an arc is directed from the vertex  $u$  to the vertex  $v$ , then it is denoted as the ordered pair  $(u, v)$ . The *degree* of a vertex  $v$  is the number of arcs incident with  $v$ . To distinguish between arcs coming into  $v$  and arcs coming out of  $v$ , we refer to the number of arcs coming into  $v$  as the *indegree* of  $v$ , while the number of arcs coming out of  $v$  is referred to as the *outdegree* of  $v$ . This is denoted as  $d^-(v)$  and  $d^+(v)$ , respectively. In evolutionary biology, directed graphs are used to represent the evolutionary history of a collection of present-day species. Vertices may represent species, individuals, or DNA sequences, while arcs represent ancestral relationships. By viewing the edges as arcs directed away from the root, rooted phylogenetic trees are examples of such digraphs.

A *directed path* in a digraph  $D$  is an alternating sequence

$$v_0, a_1, v_1, a_2, v_2, \dots, v_{k-1}, a_k, v_k$$

of vertices and arcs in which  $a_i$  is directed from  $v_{i-1}$  to  $v_i$  for all  $i$ , and no vertex or arc appears more than once. A *directed cycle* in  $D$  is a directed path in which  $v_0 = v_k$ . We say that  $D$  is *acyclic* if it contains no directed cycles. An acyclic digraph  $D$  is *rooted* if the underlying graph has no parallel edges, and there is a distinguished vertex  $\rho$  with  $d^-(\rho) = 0$  and the property that there is a directed path from  $\rho$  to every vertex of  $D$ .

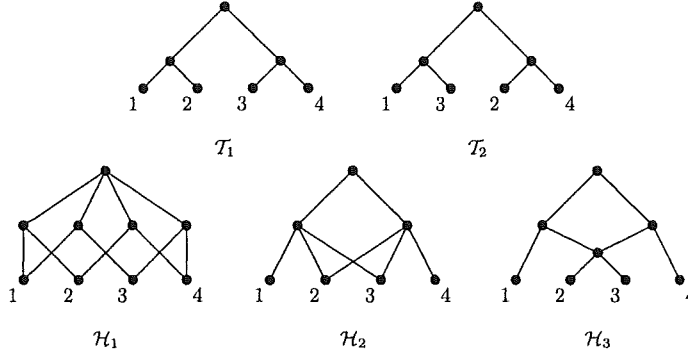


FIGURE 2. Rooted binary phylogenetic trees and hybrid phylogenies.

A *hybridization network* (on  $X$ ) is a rooted acyclic digraph with root  $\rho$  in which

- (i)  $X$  is the set of vertices of outdegree zero,
- (ii)  $d^+(\rho) \geq 2$ , and
- (iii) for all vertices  $v$  with  $d^+(v) = 1$ , we have  $d^-(v) \geq 2$ .

The set  $X$  represents a collection of taxa and is the *label set* of  $\mathcal{H}$ . For convenience, it is sometimes denoted as  $\mathcal{L}(\mathcal{H})$ . Vertices of indegree at least two represent an exchange of genetic information between their parents. Generically, we call these vertices *hybridization vertices*. In the literature, hybridization networks have been referred to as “hybrid phylogenies” (e.g., [7]) and “phylogenetic networks” (e.g., [28, 34]). The latter with the additional property that hybridization vertices have indegree exactly two. Note here that vertices with indegree more than two do not represent a simultaneous exchange of genetic information between several parents but rather an uncertainty of the exact order of speciation. To illustrate the above concepts, in Fig. 2,  $\mathcal{H}_1$ ,  $\mathcal{H}_2$ , and  $\mathcal{H}_3$  are all examples of hybridization networks in which  $X = \{1, 2, 3, 4\}$ . Here and in all other figures, it is implicit that arcs are directed downwards. Rooted phylogenetic trees are special examples of hybridization networks in which all vertices, apart from the root, have indegree 1.

To quantify the number of reticulation events, the *hybridization number* of a hybridization network  $\mathcal{H}$  with root  $\rho$  is

$$h(\mathcal{H}) = \sum_{v \neq \rho} (d^-(v) - 1).$$

Since  $d^-(v)$  is the number of parents of  $v$  and every vertex, apart from the root, has at least one parent, “ $d^-(v) - 1$ ” is the number of additional parents of  $v$ . The hybridization number of a network is at least zero. Indeed,  $h(\mathcal{H}) = 0$  if and only if  $\mathcal{H}$  is a rooted phylogenetic tree. In Fig. 2,  $h(\mathcal{H}_1) = 4$ ,  $h(\mathcal{H}_2) = 2$ , and  $h(\mathcal{H}_3) = 1$ .

Let  $\mathcal{T}$  be a rooted phylogenetic tree and let  $\mathcal{H}$  be a hybridization network. We say that  $\mathcal{H}$  *displays*  $\mathcal{T}$  if  $\mathcal{L}(\mathcal{T}) \subseteq \mathcal{L}(\mathcal{H})$  and there is a rooted subtree of  $\mathcal{H}$  that is

a refinement of  $\mathcal{T}$ . In other words,  $\mathcal{T}$  can be obtained from  $\mathcal{H}$  by first deleting a subset of the edges of  $\mathcal{H}$  and any resulting isolated vertices, and then contracting edges. For example, in Fig. 2,  $\mathcal{H}_1$  and  $\mathcal{H}_2$  both display  $\mathcal{T}_1$  and  $\mathcal{T}_2$ , while  $\mathcal{H}_3$  displays neither  $\mathcal{T}_1$  nor  $\mathcal{T}_2$ . We say that  $\mathcal{H}$  *displays* a collection  $\mathcal{P}$  of rooted phylogenetic trees if each tree in  $\mathcal{P}$  is displayed by  $\mathcal{H}$ . Furthermore, extending the definition of the hybridization number to a collection  $\mathcal{P}$  of rooted phylogenetic trees, we set

$$h(\mathcal{P}) = \min\{h(\mathcal{H}) : \mathcal{H} \text{ is a hybridization network that displays } \mathcal{P}\}.$$

If  $\mathcal{P} = \{\mathcal{T}, \mathcal{T}'\}$ , then we denote  $h(\mathcal{P})$  by  $h(\mathcal{T}, \mathcal{T}')$ .

We interpret the fundamental problem for hybridization networks for when the initial collection consists of two rooted binary phylogenetic trees as the following optimization problem:

**MINIMUM HYBRIDIZATION**

**Instance:** A finite set  $X$ , and two rooted binary phylogenetic  $X$ -trees.

**Goal:** Find a hybridization network  $\mathcal{H}$  that displays  $\mathcal{T}$  and  $\mathcal{T}'$  with minimum hybridization number.

**Measure:** The value of  $h(\mathcal{H})$ .

In Fig. 2, while  $\mathcal{H}_1$  displays  $\mathcal{T}_1$  and  $\mathcal{T}_2$ , it does not minimize the hybridization number. However, it is easily checked that  $\mathcal{H}_2$  has this property. Thus, in this case,  $h(\mathcal{T}_1, \mathcal{T}_2) = 2$ .

In its broadest sense, an instance of MINIMUM HYBRIDIZATION would consist of a collection of rooted phylogenetic trees. However, even in this simplest case when it consists of just two rooted binary phylogenetic trees, Bordewich and Semple [12] showed that MINIMUM HYBRIDIZATION is NP-hard (see Section 7). Nevertheless, there is an attractive characterization of this problem in the simplest case. This characterization provides valuable insight into the problem and is crucial to many of the results in this chapter. We describe this characterization and some of these results in the next section.

We end this section with two remarks. First, the input in the above problem could equally have been a set of sequences instead of a set of trees, in which case, instead of seeking a ‘minimal’ hybridization network, we look for a “recombination network” that has this property. A number of authors have considered this variant of the problem and we will describe it in Section 5. Second, in keeping with the terminology in the chapter written by Huson and elsewhere, we use the term “hybridization networks” as the input is unordered. In contrast, if the input is ordered in some way, as in the case of sequences, then the analogous digraphs are called “recombination networks”.

### 3. A CHARACTERIZATION OF MINIMUM HYBRIDIZATION

Historically, one of the main tools that has been used to understand and model reticulate evolution is a graph-theoretic operation called “rooted subtree prune and regraft”. Informally, this operation prunes a subtree of a rooted tree and then reattaches this subtree to another part of the tree. The use of this tool in

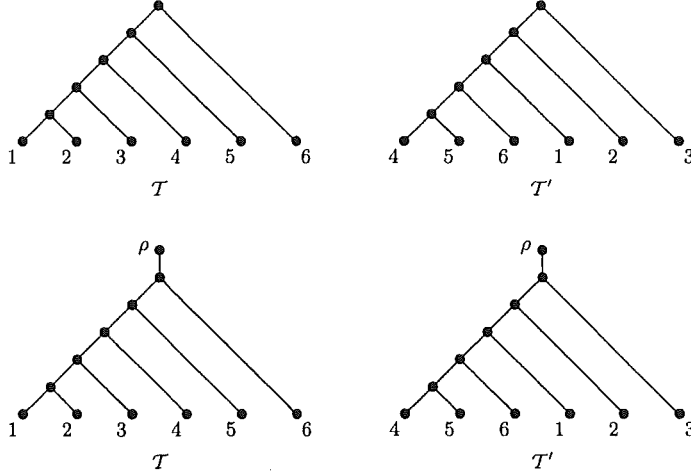


FIGURE 3. Two rooted binary phylogenetic trees  $T$  and  $T'$  without (above) and with (below) their root labelled  $\rho$ .

evolutionary biology dates back to at least 1990 [21] and has been regularly used since (for example, see [7, 30, 34, 43]) as a way to model reticulate evolution. The reason for this is that if two rooted binary phylogenetic  $X$ -trees are inconsistent, but this inconsistency can be explained with a single hybridization event, then one tree can be obtained from the other by a single rooted subtree prune and regraft operation. Indeed, given this, it is tempting to conjecture that the minimum number of hybridization events to explain the inconsistency of two rooted binary phylogenetic  $X$ -trees is equal to the minimum number of rooted subtree prune and regraft operations to transform one tree into the other. We will make this precise shortly, however, this is not the case. Nevertheless, these two minimum numbers are very closely related as they can both be characterized in terms of “agreement forests”. It is one of these characterizations that is referred to at the end of the last section.

To make the characterizations work, we regard the root of each of the two rooted binary phylogenetic  $X$ -trees  $T$  and  $T'$  in the upcoming definitions as a vertex  $\rho$  at the end of a pendant edge (called the *root edge*) adjoined to the original root. Furthermore, we regard  $\rho$  as part of the label sets of  $T$  and  $T'$ , and so  $\mathcal{L}(T) = \mathcal{L}(T') = X \cup \{\rho\}$ . To illustrate, consider the two rooted binary phylogenetic trees  $T$  and  $T'$  shown at the top of Fig. 3. In the following, we regard  $T$  and  $T'$  as shown at the bottom of Fig. 3.

**Rooted Subtree Prune and Regraft Operation.** Let  $e = \{u, v\}$  be an edge of  $T$  that is not the root edge, where  $u$  is the vertex that is on the path from the root of  $T$  to  $v$ . Let  $T'$  be the rooted binary phylogenetic tree obtained from  $T$  by deleting  $e$  and reattaching the resulting rooted subtree via a new edge,  $f$  say,

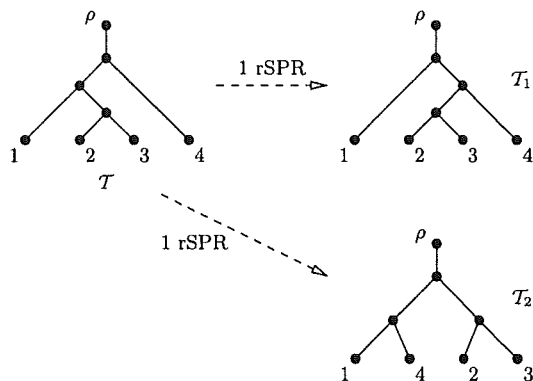


FIGURE 4. Each of  $T_1$  and  $T_2$  are obtained from  $T$  by a rooted subtree prune and regraft operation.

as follows. Create a new vertex  $u'$  that subdivides an edge of the component that contains  $\rho$  and adjoin  $f$  between  $u'$  and  $v$ , then suppress the degree-2 vertex  $u$ . We say that  $T'$  has been obtained from  $T$  by a *rooted subtree prune and regraft* (rSPR) operation. To illustrate, consider Fig. 4. Each of  $T_1$  and  $T_2$  are obtained from  $T$  by a single rSPR operation. Denoted by  $d_{\text{rSPR}}(T, T')$ , we define the rSPR *distance* between  $T$  and  $T'$  to be the minimum number of rooted subtree prune and regraft operations that is required to transform  $T$  into  $T'$ . It is well known that, for any such pair of trees, one can always obtain one tree from the other by a sequence of rSPR operations, and so this distance is well-defined. Moreover, this distance is a metric on the collection of rooted binary phylogenetic  $X$ -trees.

To explicitly highlight the connection between rooted subtree prune and regraft operations and hybridization events, consider  $T$  and  $T_1$  in Fig. 4. The evolutionary difference in the two trees can be explained by a single hybridization event; the corresponding hybridization vertex is the root of the pendant subtree that is pruned and regrafted in the rooted subtree prune and regraft operation shown in the figure.

Analogous to MINIMUM HYBRIDIZATION, we formally state the optimization problem of computing the rSPR distance between two rooted binary phylogenetic trees as follows.

#### MINIMUM RSPR

**Instance:** A finite set  $X$ , and two rooted binary phylogenetic  $X$ -trees  $T$  and  $T'$ .

**Goal:** Find a minimum length sequence of single rSPR operations that transforms  $T$  into  $T'$ .

**Measure:** The length of this sequence.



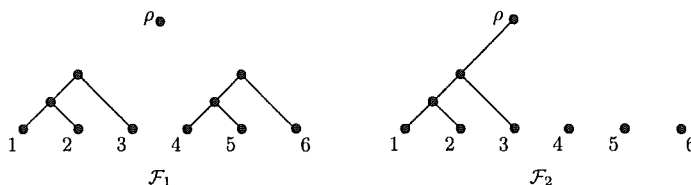


FIGURE 5. Two possible agreement forests for  $\mathcal{T}$  and  $\mathcal{T}'$  in Fig. 3.  $\mathcal{F}_1$  is a maximum-agreement forest for  $\mathcal{T}$  and  $\mathcal{T}'$ , while  $\mathcal{F}_2$  is a maximum-acyclic-agreement forest for  $\mathcal{T}$  and  $\mathcal{T}'$ .

**Agreement Forests.** An *agreement forest* for  $\mathcal{T}$  and  $\mathcal{T}'$  is a collection  $\{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_k\}$  of rooted leaf-labelled trees, where  $\mathcal{T}_\rho$  is a rooted tree whose label set  $\mathcal{L}_\rho$  contains  $\rho$  and  $\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_k$  are rooted binary phylogenetic trees with label sets  $\mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_k$ , respectively, such that the following properties are satisfied:

- (i) The label sets  $\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_k$  partition  $X \cup \{\rho\}$ .
- (ii) For each  $i \in \{\rho, 1, 2, \dots, k\}$ , we have that  $\mathcal{T}_i \cong \mathcal{T}|_{\mathcal{L}_i}$  and  $\mathcal{T}_i \cong \mathcal{T}'|_{\mathcal{L}_i}$ .
- (iii) The trees in  $\{\mathcal{T}(\mathcal{L}_i) : i \in \{\rho, 1, 2, \dots, k\}\}$  and  $\{\mathcal{T}'(\mathcal{L}_i) : i \in \{\rho, 1, 2, \dots, k\}\}$  are vertex disjoint rooted subtrees of  $\mathcal{T}$  and  $\mathcal{T}'$ , respectively.

It is easily seen that if  $\mathcal{F}$  is an agreement forest for  $\mathcal{T}$  and  $\mathcal{T}'$ , then, up to suppressing non-root vertices of degree two,  $\mathcal{F}$  can be obtained from each of  $\mathcal{T}$  and  $\mathcal{T}'$  by deleting  $|\mathcal{F}| - 1$  edges. An agreement forest for  $\mathcal{T}$  and  $\mathcal{T}'$  is a *maximum-agreement forest* if, amongst all agreement forests for  $\mathcal{T}$  and  $\mathcal{T}'$ , it has the smallest number of components, in which case we denote this value of  $k$  by  $m(\mathcal{T}, \mathcal{T}')$ . For example, two agreement forests for the two trees  $\mathcal{T}$  and  $\mathcal{T}'$  in Fig. 3 are shown in Fig. 5. It is easily checked that the smallest number of components in any such forest is 3, so  $\mathcal{F}_1$  is also a maximum-agreement forest for  $\mathcal{T}$  and  $\mathcal{T}'$ .

Intuitively, the deleted edges are those which disagree in  $\mathcal{T}$  and  $\mathcal{T}'$ , and correspond to different paths of genetic inheritance; that is hybridization events. Thus, the fewer edges deleted, the smaller the number of hybridization events. Part (i) of the following theorem by Bordewich and Semple [11] characterizes the rSPR distance between two rooted binary phylogenetic trees in terms of agreement forests.

**Theorem 3.1.** *Let  $\mathcal{T}$  and  $\mathcal{T}'$  be two rooted binary phylogenetic  $X$ -trees. Then*

- (i)  $d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}') = m(\mathcal{T}, \mathcal{T}')$ .
- (ii) *If  $\mathcal{F}$  is an agreement forest for  $\mathcal{T}$  and  $\mathcal{T}'$  of size  $k + 1$ , then there is a polynomial-time algorithm for constructing a sequence*

$$\mathcal{T} = \mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_k = \mathcal{T}'$$

*of rooted binary phylogenetic trees such that, for all  $i$ ,  $\mathcal{T}_i$  is obtained from  $\mathcal{T}_{i-1}$  by at most one rooted subtree prune and regraft operation.*

**Remarks.**

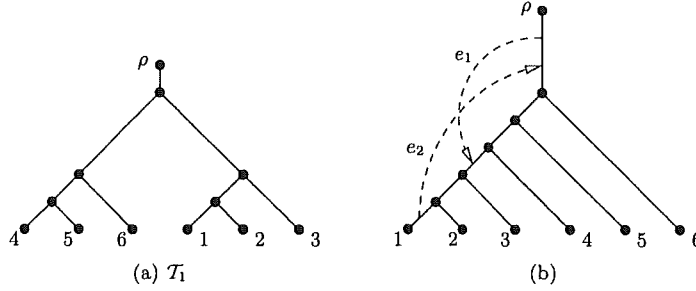


FIGURE 6. (a) The second tree in the sequence of rSPR operations that transforms  $T$  into  $T'$ . (b) The hybridization network induced by the two rSPR operations that transforms  $T$  into  $T'$ .

1. Part (ii) of Theorem 3.1 is not explicitly stated in [11]. However, it is an immediate consequence of the inductive proof of [11, Theorem 2.1]. Although we omit the proof of this result, we will describe the algorithm in (ii) later in this section.
2. For those readers familiar with the tree rearrangement operation “tree bisection and reconnection” (TBR), Allen and Steel [3] describe an analogous characterization for TBR in terms of agreement forests.
3. As we will soon see, agreement forests characterizations have been successfully used in gaining invaluable insights of various measures in phylogenetics. To provide intuition why such a characterization is useful, think how much easier it is to consider deleting edges of  $T$  and  $T'$  to obtain an agreement forest as oppose to keeping track of a sequence of rSPR operations that transforms  $T$  into  $T'$ .

Although it seems plausible that one could repeatedly used a single rooted subtree prune and regraft operation to represent a single hybridization event and thus the number of such events is equal to the number of such operations, the associated hybridization network that one builds in this process may contain a directed cycle. Such a cycle would mean that a vertex in this network inherits genetic information from its own descendants. As an example, consider the two rooted binary phylogenetic trees  $T$  and  $T'$  shown in Fig. 3. The tree  $T'$  can be obtained from  $T$  by two rSPR operations by first pruning the pendant subtree with label set  $\{1, 2, 3\}$  of  $T$  and regrafting to obtain the tree  $T_1$  in Fig. 6(a), and then pruning the pendant subtree of  $T_1$  with label set  $\{4, 5, 6\}$  and regrafting to obtain  $T'$ . If one keeps each of the edges that are cut and added in this process, one obtains the “hybridization” network shown in Fig. 6(b). Here  $e_1$  is the edge that is added in the first rSPR operation and  $e_2$  is the edge that is added in the second rSPR operation. However, by viewing the (solid) edges as arcs directed away from  $\rho$ , this network contains a directed cycle. To avoid the construction of such a cycle and, in particular, rooted subtree prune and regraft operations that cause these cycles, we extend the definition of an agreement forest to an acyclic-agreement forest.

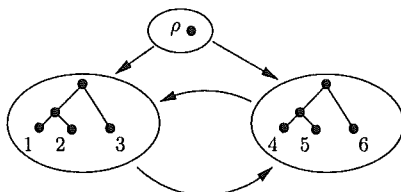


FIGURE 7. The directed graph  $G_{\mathcal{F}_1}$ , where  $\mathcal{F}_1$  is the agreement forest in Fig. 5.

Let  $\mathcal{F} = \{T_\rho, T_1, T_2, \dots, T_k\}$  be an agreement forest for  $T$  and  $T'$ . Let  $G_{\mathcal{F}}$  be the directed graph whose vertex set is  $\mathcal{F}$  and for which  $(T_i, T_j)$  is an arc precisely if  $i \neq j$  and either

- (I) the root of  $T(\mathcal{L}_i)$  is an ancestor of the root of  $T(\mathcal{L}_j)$  or
- (II) the root of  $T'(\mathcal{L}_i)$  is an ancestor of the root of  $T'(\mathcal{L}_j)$ .

Note that, as  $\mathcal{F}$  is an agreement forest, the roots of  $T(\mathcal{L}_i)$  and  $T(\mathcal{L}_j)$ , and the roots of  $T'(\mathcal{L}_i)$  and  $T'(\mathcal{L}_j)$  are not the same. We say that  $\mathcal{F}$  is *acyclic* if  $G_{\mathcal{F}}$  has no directed cycles. (Note that we have used the adjective “acyclic” here as oppose to “good” which is used in [9].) If  $\mathcal{F}$  is acyclic and it has the smallest number of components over all acyclic-agreement forests for  $T$  and  $T'$ , then  $\mathcal{F}$  is a *maximum-acyclic-agreement forest* for  $T$  and  $T'$ , in which case we denote the number  $k$  by  $m_a(T, T')$ . Observe that  $m_a(T, T') = 0$  if and only if, up to isomorphism,  $T$  and  $T'$  are identical. To illustrate these concepts, Fig. 7 shows the directed graph  $G_{\mathcal{F}_1}$  of the agreement forest  $\mathcal{F}_1$  shown in Fig. 5, where large open circles represent the vertices. Since this graph contains a directed cycle,  $\mathcal{F}_1$  is not acyclic. However, it is easily checked that  $G_{\mathcal{F}_2}$ , where  $\mathcal{F}_2$  is the agreement forest in Fig. 5 is acyclic. In fact, one can also check that this is a maximum-acyclic-agreement forest for  $T$  and  $T'$ .

Analogous to Theorem 3.1, Baroni *et al.* [9] characterized the hybridization number of two rooted binary phylogenetic trees in terms of agreement forests.

**Theorem 3.2.** *Let  $T$  and  $T'$  be two rooted binary phylogenetic  $X$ -trees. Then*

- (i)  $h(T, T') = m_a(T, T')$ .
- (ii) *If  $\mathcal{F}$  is an acyclic-agreement forest for  $T$  and  $T'$  of size  $k + 1$ , then there is a polynomial-time algorithm for constructing a hybridization network  $\mathcal{H}$  that displays  $T$  and  $T'$  with  $h(\mathcal{H}) \leq k$ .*

#### Remarks.

1. Part (ii) of Theorem 3.2 is not stated in [9], but it is an immediate consequence of its inductive proof. Like part (ii) of Theorem 3.1, we will describe the algorithm in (ii) at the end of this section.
2. In contrast to rSPR distance, the hybridization number is not a metric on the collection of rooted binary phylogenetic  $X$ -trees. To see this, consider  $T$  and  $T'$

in Fig. 3 and  $\mathcal{T}_1$  in Fig. 6. We have already noted that  $h(\mathcal{T}, \mathcal{T}') = 3$ . Furthermore, it is easily checked that  $h(\mathcal{T}, \mathcal{T}_1) = h(\mathcal{T}_1, \mathcal{T}') = 1$ , and so the hybridization number does not satisfy the triangle inequality.

3. If one is only interested in the number of hybridization vertices (and not what each such vertex contributes to the hybridization number), then Theorem 3.2 is easily generalized to an arbitrary size collection of rooted binary phylogenetic  $X$ -trees. Here the notion of an agreement forest is extended in the obvious way.

Since every acyclic-agreement forest for two rooted binary phylogenetic  $X$ -trees  $\mathcal{T}$  and  $\mathcal{T}'$  is an (ordinary) agreement forest for  $\mathcal{T}$  and  $\mathcal{T}'$ , it follows from Theorems 3.1 and 3.2 that

$$(1) \quad d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}') \leq h(\mathcal{T}, \mathcal{T}').$$

The fact that this inequality can be strict has been pointed out several times in the literature including [9, 22, 45]. An interesting question is just how strict? We consider this question in Section 3.1.

**3.1. Comparing  $d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}')$  and  $h(\mathcal{T}, \mathcal{T}')$ .** Two natural questions arise from the inequality in (1).

- (i) Whenever  $d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}') = 1$ , we have that  $h(\mathcal{T}, \mathcal{T}') = 1$ , and so  $d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}')$  provides a sharp lower bound for  $h(\mathcal{T}, \mathcal{T}')$ . Can we find a sharp upper bound for  $h(\mathcal{T}, \mathcal{T}')$ ?
- (ii) We have already seen that inequality (1) can be strict, so how large can the difference between  $d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}')$  and  $h(\mathcal{T}, \mathcal{T}')$  be?

Consider (i). Regardless of the topology of  $\mathcal{T}$  and  $\mathcal{T}'$ , if  $X = \{x_1, x_2, \dots, x_n\}$ , then, as the forest consisting of  $\mathcal{T} \setminus \{\rho, x_1, x_2\}$  and isolated vertices  $x_3, x_4, \dots, x_n$  is an acyclic-agreement forest for  $\mathcal{T}$  and  $\mathcal{T}'$ ,

$$h(\mathcal{T}, \mathcal{T}') \leq n - 2.$$

Using Theorem 3.2, Baroni *et al.* [9] showed that this upper bound is sharp. In particular, if  $\mathcal{T}$  and  $\mathcal{T}'$  are the two rooted caterpillars  $(x_1, x_2, \dots, x_n)$  and  $(x_n, x_{n-1}, \dots, x_1)$ , then  $h(\mathcal{T}, \mathcal{T}') = n - 2$ . In the same paper [9] and using Theorems 3.1 and 3.2, the authors also establish the following theorem.

**Theorem 3.3.** *For all  $n \geq 4$ , there are rooted binary phylogenetic trees  $\mathcal{T}_1$ ,  $\mathcal{T}_2$ , and  $\mathcal{T}_3$  on  $n$  leaves such that*

$$\frac{h(\mathcal{T}_1, \mathcal{T}_2)}{d_{\text{rSPR}}(\mathcal{T}_1, \mathcal{T}_2)} = \frac{1}{2} \left\lfloor \frac{n}{2} \right\rfloor$$

and

$$h(\mathcal{T}_1, \mathcal{T}_3) - d_{\text{rSPR}}(\mathcal{T}_1, \mathcal{T}_3) = n - 2 \lfloor \sqrt{n} \rfloor - c,$$

where  $c = 0$  if  $n$  is a square,  $c = 1$  if  $1 \leq n - \lfloor \sqrt{n} \rfloor^2 < \sqrt{n}$ , and  $c = 2$  otherwise.

Explicit examples of rooted binary phylogenetic trees that attain the equalities in Theorem 3.3 are given in [9]. For example, let  $\mathcal{T}_1$  be the rooted caterpillar tree

$(x_1, x_2, \dots, x_{100})$ . Let  $\mathcal{T}_2$  and  $\mathcal{T}_3$  be the rooted caterpillar trees on  $\{x_1, x_2, \dots, x_{100}\}$  whose orderings on their leaf sets are

$$(x_{51}, x_{52}, \dots, x_{100}, x_1, x_2, \dots, x_{50})$$

and

$$(x_{91}, x_{92}, \dots, x_{100}, x_{81}, x_{82}, \dots, x_{90}, x_{71}, \dots, x_{19}, x_{20}, x_1, x_2, \dots, x_{10}),$$

respectively. Then

$$\frac{h(\mathcal{T}_1, \mathcal{T}_2)}{d_{\text{rSPR}}(\mathcal{T}_1, \mathcal{T}_2)} = \frac{1}{2} \left\lfloor \frac{100}{2} \right\rfloor = 25$$

and

$$h(\mathcal{T}_1, \mathcal{T}_3) - d_{\text{rSPR}}(\mathcal{T}_1, \mathcal{T}_3) = 100 - 2\lfloor\sqrt{100}\rfloor - 0 = 80$$

An interesting question is determine whether the ratio or difference given in Theorem 3.3 is the best possible.

The answers to (i) and (ii) in [9] both rely on Theorems 3.1 and 3.2. It seems unlikely that, without such characterizations, such results could have been attained as easily. Further applications of these theorems are given in Section 4.

**3.2. Algorithms for constructing rSPR sequences and hybridization networks from agreement forests.** Let  $\mathcal{F}$  be an agreement forest for two rooted binary phylogenetic  $X$ -trees  $\mathcal{T}$  and  $\mathcal{T}'$ . Note that  $\mathcal{F}$  need not be optimal. The first algorithm `RSPRSEQUENCE` constructs a sequence of rooted binary phylogenetic trees beginning with  $\mathcal{T}$  and ending with  $\mathcal{T}'$  with the property that each tree in the sequence is obtained from its predecessor by a single rSPR operation. Provided  $\mathcal{F}$  is acyclic, the second algorithm `HYBRIDNETWORK` constructs a hybridization network  $\mathcal{H}$  that displays  $\mathcal{T}$  and  $\mathcal{T}'$  with  $h(\mathcal{H}) \leq |\mathcal{F}| - 1$ . Each algorithm is an immediate consequence of the inductive proofs of Theorems 3.1 and 3.2 in [11] and [9], respectively.

**Algorithm:** `RSPRSEQUENCE`( $\mathcal{F}$ )

**Input:** An agreement forest  $\mathcal{F}$  of size  $k + 1$  of two rooted binary phylogenetic  $X$ -trees  $\mathcal{T}$  and  $\mathcal{T}'$ .

**Output:** A sequence  $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_k$  of rooted binary phylogenetic  $X$ -trees with the property that  $\mathcal{T}_0 = \mathcal{T}$ ,  $\mathcal{T}_k = \mathcal{T}'$ , and, for all  $i$ , either  $\mathcal{T}_i$  is obtained from  $\mathcal{T}_{i-1}$  by a single rSPR operation or  $\mathcal{T}_i \cong \mathcal{T}_{i-1}$ .

1. Set  $\mathcal{T} = \mathcal{T}_0$ ,  $\mathcal{F} = \mathcal{F}_0$ , and  $i = 1$ .
2. Find a tree  $\mathcal{S}_i$  in  $\mathcal{F}_{i-1}$  such that  $\mathcal{S}_i$  can be pruned from the rest of  $\mathcal{T}_{i-1}$  by deleting a single edge.
3. In  $\mathcal{T}'$ , find the first subtree  $\mathcal{T}'(\mathcal{L}(\mathcal{S}_j))$  corresponding to a tree  $\mathcal{S}_j$  in  $\mathcal{F}_{i-1}$  that is met on the path from the root of  $\mathcal{T}'(\mathcal{L}(\mathcal{S}_i))$  to  $\rho$ .
4. Set  $\mathcal{T}_i$  to be a tree that is obtained from  $\mathcal{T}_{i-1}$  by pruning  $\mathcal{S}_i$  and regrafting it so that  $\mathcal{T}_i$  restricted to  $\mathcal{L}(\mathcal{S}_i) \cup \mathcal{L}(\mathcal{S}_j)$  is isomorphic to  $\mathcal{T}'$  restricted to  $\mathcal{L}(\mathcal{S}_i) \cup \mathcal{L}(\mathcal{S}_j)$ .
5. Set  $\mathcal{F}_i$  to be the forest obtained from  $\mathcal{F}_{i-1}$  by replacing  $\mathcal{S}_i$  and  $\mathcal{S}_j$  with  $\mathcal{T}'$  restricted to  $\mathcal{L}(\mathcal{S}_i) \cup \mathcal{L}(\mathcal{S}_j)$ .
6. If  $i = k$  halt; otherwise, increment  $i$  by 1 and return to Step 2.

**Remarks.** The following comments may help the reader.

1. Step 2 is well-defined as there is always at least one tree that has this property.
2. In Step 3, the choice for  $\mathcal{S}_j$  is unique because of (iii) in the definition of an agreement forest.
3. In Step 4,  $\mathcal{F}_i$  is an agreement forest for  $\mathcal{T}_i$  and  $\mathcal{T}'$ .

Before stating HYBRIDNETWORK, we need an additional concept. A simple, fast, and well-known way of deciding whether a directed graph  $G$  is acyclic is as follows. Find a vertex,  $v_1$  say, of  $G$  that has indegree 0. If there is no such vertex, then  $G$  contains a directed cycle and so  $G$  is acyclic. Otherwise, delete  $v_1$  (and its incident arcs) from  $G$  and find a vertex,  $v_2$  say, of  $G$  that has indegree 0. Again, if there is no such vertex, then  $G$  is not acyclic, otherwise delete  $v_2$  from this last digraph and continue in this way. Eventually, we either decide that  $G$  is not acyclic or obtain an ordering  $v_1, v_2, \dots, v_n$  of the vertex set of  $G$  such that, for all  $i$ , the vertex  $v_i$  has indegree 0 in the graph obtained from  $G$  by deleting the vertices  $v_1, v_2, \dots, v_{i-1}$ . Such an ordering is called an *acyclic ordering* of  $G$  and it implies that  $G$  is acyclic.

**Algorithm:** HYBRIDNETWORK( $\mathcal{F}$ )

**Input:** An acyclic-agreement forest  $\mathcal{F}$  of size  $k+1$  of two rooted binary phylogenetic  $X$ -trees  $\mathcal{T}$  and  $\mathcal{T}'$ .

**Output:** A hybridization network  $\mathcal{H}$  that displays  $\mathcal{T}$  and  $\mathcal{T}'$  with  $h(\mathcal{H}) \leq k$ .

1. Find an acyclic ordering,  $\mathcal{S}_\rho, \mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_k$  say, of  $G_{\mathcal{F}}$ .
2. Set  $\mathcal{H}_0 = \mathcal{S}_\rho$  and set  $i = 1$ .
3. Attach  $\mathcal{S}_i$  to  $\mathcal{H}_{i-1}$  via two new arcs. Each arc joins the root of  $\mathcal{S}_i$  to some (possibly distinct) arc of  $\mathcal{H}_{i-1}$  and is directed towards the root of  $\mathcal{S}_i$ . These arcs are added so that the resulting network displays both  $\mathcal{T}$  restricted to  $\mathcal{L}(\mathcal{H}_{i-1}) \cup \mathcal{L}(\mathcal{S}_i)$  and  $\mathcal{T}'$  restricted to  $\mathcal{L}(\mathcal{H}_{i-1}) \cup \mathcal{L}(\mathcal{S}_i)$ .  
Set  $\mathcal{H}_i$  to be the resulting network and return  $\mathcal{H}_i$  if  $i = k$ .
4. Increment  $i$  by 1 and return to Step 3.

**Remark.** In Step 3 of the algorithm, it may be possible that only one new edge is required. This implies that  $\mathcal{F}$  is not maximum and that a new acyclic-agreement forest for  $\mathcal{T}$  and  $\mathcal{T}'$  can be obtained by attaching one component  $\mathcal{S}$  of  $\mathcal{F}$  to another via an edge directed towards the root of  $\mathcal{S}$ .

#### 4. APPLICATIONS OF AGREEMENT FORESTS

For two rooted binary phylogenetic trees  $\mathcal{T}$  and  $\mathcal{T}'$ , agreement forests are a particularly useful tool for analyzing the values  $d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}')$  and  $h(\mathcal{T}, \mathcal{T}')$ . In this section, we consider ways that agreement forests can be used for this analysis and the resulting algorithmic implications, while in Section 7 we see that this tool provides invaluable leverage in understanding the computation complexity of finding these values.

As we formally state in Section 7, both MINIMUM RSPR and MINIMUM HYBRIDIZATION are NP-hard problems. Nevertheless, they are both susceptible to

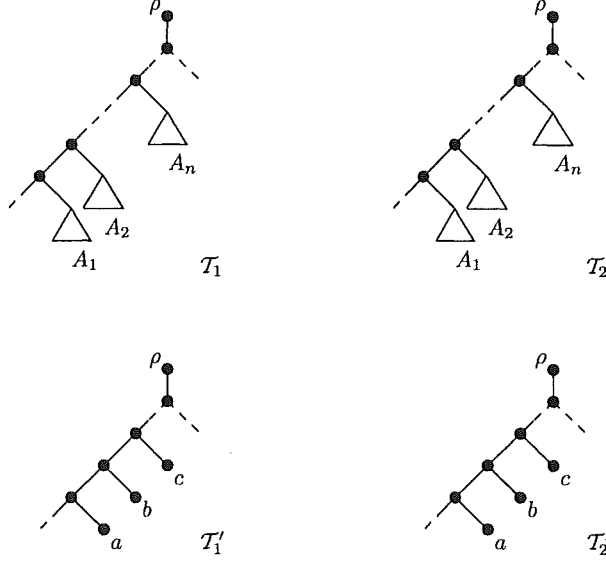


FIGURE 8. Applying Rule 2 to two rooted binary phylogenetic trees  $T_1$  and  $T_2$ .

approaches that effectively reduce the size of the problem instance. Interestingly, these approaches are different and it appears that they are unique to the particular problem. For MINIMUM RSPR, we reduce the size of the problem instance while preserving the rooted subtree prune and regraft distance, while, for MINIMUM HYBRIDIZATION, we use a divide-and-conquer type approach, that is we break the problem into a number of smaller problems. To avoid some repetition, the proofs of the first four results in this section rely on either Theorem 3.1 or Theorem 3.2.

For MINIMUM RSPR, consider the following two reduction rules:

- Rule 1. Replace a pendant subtree that occurs identically in both trees by a single leaf with a new label.
- Rule 2. Replace a chain of at least three pendant subtrees that occur identically and with the same orientation relative to the root in both trees by three new leaves with new labels correctly orientated to preserve the direction of the chain.

Rule 2 is illustrated in Fig. 8, where  $A_1, A_2, \dots, A_n$  is the chain of pendant subtrees common to both  $T_1$  and  $T_2$ , and  $a, b$ , and  $c$  are the three new leaf labels orientated appropriately.

The following theorem is due to Bordewich and Semple [11].

**Theorem 4.1.** *Let  $T_1$  and  $T_2$  be two rooted binary phylogenetic  $X$ -trees, and let  $T'_1$  and  $T'_2$  be the two rooted binary phylogenetic  $X'$ -trees obtained from  $T_1$  and  $T_2$ , respectively, by applying either Rule 1 or Rule 2. Then*

$$d_{\text{rSPR}}(T_1, T_2) = d_{\text{rSPR}}(T'_1, T'_2).$$

The proof of Theorem 4.1 relies on Theorem 3.1 and is the basis of showing that MINIMUM rSPR is fixed-parameter tractable in  $d_{\text{rSPR}}(T_1, T_2)$ . Intuitively, this simply means that if the rSPR distance is small, it may be possible to efficiently compute this distance even if  $X$  is large. The reason for this is that, for small rSPR distance, one would expect the problem instance to be significantly reduced by repeatedly applying Rules 1 and 2. Note that, by Theorem 4.1, such repeated applications preserve the rSPR distance. For further details, see Section 7.

For MINIMUM HYBRIDIZATION, we have the following theorem due to Baroni [6] (also see Baroni *et al.* [8]), which provides a divide-and-conquer type approach to the problem.

**Theorem 4.2.** *Let  $T$  and  $T'$  be two rooted binary phylogenetic  $X$ -trees, and suppose that  $A \subset X$  is a cluster of both  $T$  and  $T'$ . Then*

$$h(T, T') = h(T|A, T'|A) + h(T_a, T'_a),$$

where  $T_a$  and  $T'_a$  are obtained from  $T$  and  $T'$ , respectively, by replacing the pendant subtrees  $T(A)$  and  $T'(A)$  with a single new leaf labelled  $a$ . Furthermore, if  $\mathcal{H}_a$  is a hybridization network that displays  $T_a$  and  $T'_a$  with  $h(\mathcal{H}_a) = h(T_a, T'_a)$  and  $\mathcal{H}_A$  is a hybridization network that displays  $T|A$  and  $T'|A$  with  $h(\mathcal{H}_A) = h(T|A, T'|A)$ , then the hybridization network obtained from  $\mathcal{H}_a$  by identifying the root of  $\mathcal{H}_A$  with  $a$  displays  $T$  and  $T'$ , and has hybridization number  $h(T, T')$ .

We will discuss the obvious divide-and-conquer algorithm resulting from Theorem 4.2 in Section 4.1.

Recalling that if, up to isomorphism, two rooted binary phylogenetic trees are identical, then their hybridization number is 0, we get the following corollary as an immediate consequence of Theorem 4.2.

**Corollary 4.3.** *Let  $T_1$  and  $T_2$  be two rooted binary phylogenetic  $X$ -trees, and let  $T'_1$  and  $T'_2$  be the two rooted binary phylogenetic  $X'$ -trees obtained from  $T_1$  and  $T_2$ , respectively, by applying Rule 1. Then*

$$h(T_1, T_2) = h(T'_1, T'_2).$$

Curiously, despite Corollary 4.3, Rule 2 does not preserve the hybridization number of two rooted binary phylogenetic trees. We illustrate with a simple example. Let  $T_1$  and  $T_2$  be the rooted caterpillar trees

$$(b_1, b_2, b_3, b_4, b_5, b_6, a_1, a_2, a_3, a_4)$$



and

$$(b_1, a_1, a_2, a_3, a_4, b_2, b_3, b_4, b_5, b_6),$$

respectively. Let  $T'_1$  and  $T'_2$  be the rooted caterpillar trees obtained from  $T_1$  and  $T_2$ , respectively, by applying Rule 2 to the chain of pendant subtrees corresponding to the labels  $a_1, a_2, a_3, a_4$ . Let  $a$ ,  $b$ , and  $c$  denote the resulting new leaves. Thus  $T'_1$  and  $T'_2$  are the rooted caterpillar trees

$$(b_1, b_2, b_3, b_4, b_5, b_6, a, b, c)$$

and

$$(b_1, a, b, c, b_2, b_3, b_4, b_5, b_6),$$

respectively. First observe that the agreement forest  $\mathcal{F}$  of  $T_1$  and  $T_2$  for which the partition of  $X \cup \{\rho\}$  induced by the label sets of its trees is

$$\{\{b_1, b_2, b_3, b_4, b_5, b_6, \rho\}, \{a_1\}, \{a_2\}, \{a_3\}, \{a_4\}\}$$

is acyclic. Thus the number of components of a maximum-acyclic-agreement forest of  $T_1$  and  $T_2$  is at most 5. We next show that this number is exactly 5 and that  $\mathcal{F}$  is the unique maximum-acyclic agreement forest for  $T_1$  and  $T_2$ . Let  $\mathcal{F}'$  be a maximum-acyclic-agreement forest for  $T_1$  and  $T_2$ . If  $b_j \in \mathcal{L}_\rho$  for some  $j$ , then, by the maximality of  $\mathcal{F}'$ ,  $\{a_1\}, \{a_2\}, \{a_3\}, \{a_4\}$  are label sets of  $\mathcal{F}'$  and so, as  $\mathcal{F}'$  is maximum,  $\mathcal{F}' = \mathcal{F}$ . Furthermore, if  $a_i \in \mathcal{L}_\rho$  for some  $i$ , then  $\{b_2\}, \{b_3\}, \{b_4\}, \{b_5\}, \{b_6\}$  are label sets of  $\mathcal{F}'$  and so  $|\mathcal{F}'| \geq 6$ ; a contradiction to maximality. Thus  $\{\rho\}$  is a label set of  $\mathcal{F}'$ , in particular  $\mathcal{L}_\rho \cap X$  is empty. But, because of the necessity of being acyclic,  $\mathcal{L}_\rho \cap X$  is non-empty in any maximum-acyclic-agreement forest for  $T_1$  and  $T_2$  [9]. This last contradiction shows that  $\mathcal{F}$  is the unique maximum-acyclic-agreement forest for  $T_1$  and  $T_2$ . Using similar arguments, the unique maximum-acyclic-agreement forest for  $T'_1$  and  $T'_2$  is the forest for which the partition of  $X \cup \{\rho\}$  induced by the label sets of its trees is

$$\{\{b_1, b_2, b_3, b_4, b_5, b_6, \rho\}, \{a\}, \{b\}, \{c\}\}.$$

But then  $h(T_1, T_2) = 4$ , while  $h(T'_1, T'_2) = 3$ . Thus Rule 2 does not preserve the hybridization number of two trees. The main point of the argument above is that, unlike the situation for (ordinary) agreement forests, there is no maximum-acyclic-agreement forest that contains a tree whose label set contains  $\{a_1, a_2, a_3, a_4\}$ , the union of the label sets of the chain of pendant subtrees that are replaced by the three new leaves.

In comparison to the last paragraph, the rSPR distance only satisfies a weaker version of Theorem 4.2. In particular, we have the following result [11].

**Proposition 4.4.** *Let  $T$  and  $T'$  be two rooted binary phylogenetic  $X$ -trees, and suppose that  $A \subset X$  is a cluster of both  $T$  and  $T'$ . Then*

$$d_{\text{rSPR}}(T, T') \leq d_{\text{rSPR}}(T|A, T'|A) + d_{\text{rSPR}}(T_a, T'_a) \leq d_{\text{rSPR}}(T, T') + 1,$$

where  $T_a$  and  $T'_a$  are obtained from  $T$  and  $T'$ , respectively, by replacing the pendant subtrees  $T(A)$  and  $T'(A)$  with a single new leaf labelled  $a$ . Moreover, these bounds are sharp.

To see that the first bound in Proposition 4.4 is sharp, simply choose  $T$  and  $T'$  so that  $d_{\text{rSPR}}(T, T') = 1$ , and choose  $A$  to be the cluster of the pendant subtree that is pruned. For the sharpness of the second bound, choose  $T$  and  $T'$  to be the

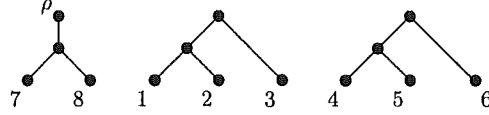


FIGURE 9. Illustrating strict inequality in Proposition 4.4.

rooted caterpillar trees  $(1, 2, 3, 4, 5, 6, 7, 8)$  and  $(4, 5, 6, 1, 2, 3, 8, 7)$ , and choose  $A$  to be the common cluster  $\{1, 2, 3, 4, 5, 6\}$ . Then  $d_{\text{rSPR}}(\mathcal{T}_a, \mathcal{T}'_a) = 1$  and, as we have seen previously,  $d_{\text{rSPR}}(\mathcal{T}|A, \mathcal{T}'|A) = 2$ , so

$$d_{\text{rSPR}}(\mathcal{T}|A, \mathcal{T}'|A) + d_{\text{rSPR}}(\mathcal{T}_a, \mathcal{T}'_a) = 3.$$

But the forest shown in Fig. 9 is an agreement forest for  $\mathcal{T}$  and  $\mathcal{T}'$ , and therefore  $d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}') \leq 2$ .

In the rest of this section, we describe two applications of Theorem 4.2.

**4.1. A simple divide-and-conquer algorithm for MINIMUM HYBRIDIZATION.** Proposition 4.2 and Corollary 4.3 provides us with the following simple divide-and-conquer approach to MINIMUM HYBRIDIZATION that is somewhat better than the naive approach of exhaustively searching for edges in  $\mathcal{T}$  (or  $\mathcal{T}'$ ) whose deletion results in an acyclic-agreement forest. This exact algorithm initially applies Rule 1 to  $\mathcal{T}$  and  $\mathcal{T}'$  as much as possible, and then locates the smallest pendant subtrees,  $\mathcal{S}$  and  $\mathcal{S}'$  say, in the resulting trees whose leaf sets are equal. These pendant subtrees localize conflicting signals in the evolutionary history of these parts of  $\mathcal{T}$  and  $\mathcal{T}'$  (see Proposition 4.5 below). The algorithm finds a maximum-acyclic-agreement forest for these pendant subtrees  $\mathcal{S}$  and  $\mathcal{S}'$ , and then repeats this process for the rooted binary phylogenetic trees obtained from  $\mathcal{S}$  and  $\mathcal{S}'$  by replacing the pendant subtrees with a single new vertex. Summing the hybridization number  $h(\mathcal{S}, \mathcal{S}')$  at each iteration gives  $h(\mathcal{T}, \mathcal{T}')$ .

**Algorithm:** HYBRIDNUMBER( $\{\mathcal{T}, \mathcal{T}'\}$ )

**Input:** Two rooted binary phylogenetic  $X$ -trees  $\mathcal{T}$  and  $\mathcal{T}'$

**Output:** The value of  $h(\mathcal{T}, \mathcal{T}')$ .

1. Set  $\mathcal{T}_0 = \mathcal{T}$  and  $\mathcal{T}'_0 = \mathcal{T}'$ , and set  $i = 1$
2. Repeatedly apply Rule 1 to  $\mathcal{T}_{i-1}$  and  $\mathcal{T}'_{i-1}$  until the rule can no longer be applied, and set  $\mathcal{S}_{i-1}$  and  $\mathcal{S}'_{i-1}$  to be the resulting rooted binary phylogenetic trees, respectively. If each of  $\mathcal{S}_{i-1}$  and  $\mathcal{S}'_{i-1}$  consist of a single vertex, then go to Step 7.
3. Find a minimal cluster  $W_{i-1}$  in  $\mathcal{C}(\mathcal{S}_{i-1}) \cap \mathcal{C}(\mathcal{S}'_{i-1})$  of size at least two.
4. Find a maximum-acyclic-agreement forest  $\mathcal{F}_{i-1}$  for  $\mathcal{S}_{i-1}|W_{i-1}$  and  $\mathcal{S}'_{i-1}|W_{i-1}$ .
5. Set  $\mathcal{T}_i$  and  $\mathcal{T}'_i$  to be the rooted binary phylogenetic trees obtained from  $\mathcal{S}_{i-1}$  and  $\mathcal{S}'_{i-1}$ , respectively, by replacing  $\mathcal{S}_{i-1}|W_{i-1}$  and  $\mathcal{S}'_{i-1}|W_{i-1}$  with a single new vertex  $w_{i-1}$ .
6. Increment  $i$  by 1 and return to Step 2.
7. Output the sum  $|\mathcal{F}_0| - 1 + |\mathcal{F}_1| - 1 + \dots + |\mathcal{F}_{i-1}| - 1$ .

**Remark.** A naive approach to Step 4 is to exhaustively delete edges from one of the trees,  $\mathcal{T}$  say, and then see if the resulting forest is an acyclic-agreement forest for  $\mathcal{T}$  and  $\mathcal{T}'$ . Furthermore, observe that, if one ignores the task of finding a maximum-acyclic-agreement forest in Step 4, then HYBRIDNUMBER provides a fast lower bound for  $h(\mathcal{T}, \mathcal{T}')$ . In particular, the number of iterations of the algorithm.

Clearly, Step 4 is the computationally most expensive part of the algorithm. However, although there is no theoretical foundations for the complexity of this algorithm, it will work well in practice provided it breaks the problem into a number of isolated parts for which the associated hybridization number is relatively small. To see whether this proviso is realistic or not, Linz *et al.* [29] carried out an experimental analysis of HYBRIDNUMBER on a grass data set that had previously been considered by Schmidt [36]. The analysis involved running the algorithm on pairs of trees with up to 40 taxa. With regards to the running time, the algorithm performed well in many instances. For example, one pair of trees on 30 taxa has a hybridization number of 8, yet the algorithm returned the answer in 111 seconds. As well as MINIMUM HYBRIDIZATION, the paper also analyzes the fixed-parameter algorithm for MINIMUM RSPR mention earlier in this section on the same data set. For further details, see [29].

As an aside, the subtrees of  $\mathcal{T}$  and  $\mathcal{T}'$  considered in Step 4 are exactly the parts of  $\mathcal{T}$  and  $\mathcal{T}'$  that conflict. More precisely, we have the following proposition whose proof is omitted. The *cluster incompatibility graph* of two rooted binary phylogenetic  $X$ -trees  $\mathcal{T}$  and  $\mathcal{T}'$  is the graph whose vertex set is  $\mathcal{C}(\mathcal{T}) \cup \mathcal{C}(\mathcal{T}')$ , and where an edge joins two vertices  $A$  and  $B$  precisely if there does not exist a rooted binary phylogenetic tree whose cluster set contains both  $A$  and  $B$ . Equivalently, this means that  $A \cap B \not\subseteq \{\emptyset, A, B\}$  (see [37]).

**Proposition 4.5.** *Let  $\mathcal{T}$  and  $\mathcal{T}'$  be two rooted binary phylogenetic  $X$ -trees, and let  $S$  and  $S'$  be subtrees of  $\mathcal{T}$  and  $\mathcal{T}'$ , respectively, considered in Step 4 of HYBRIDNUMBER. Let  $u$  and  $u'$  be the degree-2 root vertices of  $S$  and  $S'$ . Then*

$$\begin{aligned} &\{C_{\mathcal{T}}(v) : v \in V(S) - \{u\} \text{ and } d_S^+(v) \neq 0\} \\ &\cup \{C_{\mathcal{T}'}(v) : v \in V(S') - \{u'\} \text{ and } d_{S'}^+(v) \neq 0\} \end{aligned}$$

*is the vertex set of a component of the cluster incompatibility graph of  $\mathcal{T}$  and  $\mathcal{T}'$ . Moreover, the isolated vertices of this graph are precisely the clusters common to both  $\mathcal{T}$  and  $\mathcal{T}'$ .*

We end this subsection with two further comments. Firstly, Nakhleh *et al.* [33] describe a polynomial-time heuristic for finding  $h(\mathcal{T}, \mathcal{T}')$  that is based on an agreement forest type approach. In this heuristic, they obtain a certain “agreement” forest by repeatedly finding a maximum-agreement subtree of two trees to decompose  $\mathcal{T}$  and  $\mathcal{T}'$ . For further details and the associated reconstruction algorithm, see [33]. Secondly, although we have not included the details here, it is straightforward to construct a hybridization network associated with HYBRIDNUMBER by combining our earlier algorithm HYBRIDNETWORK with the second part of Theorem 4.2. However, it is important to note that such a network is not necessarily unique. Typically, there will be a number of possibilities.

**4.2. Galled-tree hybridization networks.** Whenever one is confronted with an NP-hard problem, a natural consideration is to see if there exists a polynomial-time algorithm for special instances of the problem that are still meaningful. In this subsection, we describe one particular instance that has been very successful in this regard.

Ignoring the directions of the arcs, a *galled-tree hybridization network* is a hybridization network in which every vertex is in at most one cycle. This means that, for every pair of cycles, their vertex sets (and thus arc sets) are disjoint. For ease of reading, we will refer to such networks as *gall-tree networks*. In keeping with the terminology in the literature, a cycle in a gall-tree network is called a *gall*. First studied in [46], gall-tree networks have been subsequently studied both in the hybridization and recombination settings (see Section 5 for details on the latter setting). These include algorithmic studies [17, 18, 28, 34, 41] and enumeration studies [38]. The motivation for their study is that, as hybridization events are relatively rare, one may expect such events to be isolated in which case conflicts in the initial collection of phylogenetic trees could be explained by a gall-tree network.

Let  $\mathcal{T}$  and  $\mathcal{T}'$  be two rooted binary phylogenetic  $X$ -trees, and let  $|X| = n$ . Nakhleh *et al.* [34] describe an  $O(mn)$  algorithm for deciding if there exists a gall-tree network that displays  $\mathcal{T}$  and  $\mathcal{T}'$ , and then constructs such a minimal network, where  $m$  is the hybridization number of this network. Note that there is a proviso on the network that they construct, in particular, it is minimal with respect to all other gall-tree networks that display  $\mathcal{T}$  and  $\mathcal{T}'$ . However, this proviso is not necessary because of the following proposition.

**Proposition 4.6.** *Let  $\mathcal{T}$  and  $\mathcal{T}'$  be two rooted binary phylogenetic  $X$ -trees, and suppose that there is a galled-tree network that displays  $\mathcal{T}$  and  $\mathcal{T}'$ . Suppose that the smallest number of hybridization vertices in such a network is  $m$ . Then  $h(\mathcal{T}, \mathcal{T}') = m$ .*

Before proving Proposition 4.6, we remark that the algorithm in [34] is essentially equivalent to combining HYBRIDNUMBER and HYBRIDNETWORK, and so one could establish the proposition as a consequence of these algorithms. However, we prove it directly using Theorem 4.2.

*Proof of Proposition 4.6.* The proof is by induction on  $m$ . If  $m = 0$ , then  $\mathcal{T}$  and  $\mathcal{T}'$  are isomorphic, so  $h(\mathcal{T}, \mathcal{T}') = 0$  and the theorem holds. Now suppose that  $m = k+1$  for some  $k \geq 0$  and that the theorem holds whenever the smallest number of hybridization vertices in a galled-tree network that displays the two input trees is at most  $k$ .

Let  $\mathcal{H}$  be galled-tree network that displays  $\mathcal{T}$  and  $\mathcal{T}'$ , and has the smallest number of hybridization vertices amongst all such networks. Because of the minimality condition, we may assume that each hybridization vertex has indegree 2. Referring to the unique vertex of a gall that is closest to the root than any other vertex of the gall as the coalescent vertex of the gall, let  $w$  be the coalescent vertex of a gall  $Q$  in  $\mathcal{H}$  such that there is no directed path in  $\mathcal{H}$  from  $w$  to another vertex that is the coalescent vertex of a gall in  $\mathcal{H}$ . Before continuing, we make two observations:

- (i) The subset  $W$  of  $X$  whose elements can be reached from  $w$  via a directed path is a cluster of both  $\mathcal{T}$  and  $\mathcal{T}'$ .
- (ii) The subtree of  $\mathcal{T}$  induced by  $W$  can be obtained from the subnetwork of  $\mathcal{H}$  that consists of all vertices and arcs that lie on a directed path from  $w$  by deleting one of the incoming arcs of the hybridization vertex in  $Q$ . Similarly, the subtree of  $\mathcal{T}'$  induced by  $W$  can be obtained by deleting the other incoming arc of the hybridization vertex in  $Q$ .

Let  $\mathcal{T}_w$  and  $\mathcal{T}'_w$  be the rooted binary phylogenetic trees obtained from  $\mathcal{T}$  and  $\mathcal{T}'$ , respectively, by replacing the subtrees  $\mathcal{T}|W$  and  $\mathcal{T}'|W$  with a single vertex labelled  $w$ , where  $w \notin X$ . By Theorem 4.2,

$$h(\mathcal{T}, \mathcal{T}') = h(\mathcal{T}|W, \mathcal{T}'|W) + h(\mathcal{T}_w, \mathcal{T}'_w).$$

Since  $\mathcal{T}|W$  is not isomorphic to  $\mathcal{T}'|W$ , we have that  $h(\mathcal{T}|W, \mathcal{T}'|W) \geq 1$ . But, by (ii),  $h(\mathcal{T}|W, \mathcal{T}'|W) \leq 1$  and therefore  $h(\mathcal{T}|W, \mathcal{T}'|W) = 1$ . Consider  $h(\mathcal{T}_w, \mathcal{T}'_w)$ . Let  $\mathcal{H}_w$  denote the gall-tree network obtained from  $\mathcal{H}$  by deleting each of the vertices that lie on a directed path from  $w$  except  $w$  itself. Since  $\mathcal{H}$  displays  $\mathcal{T}$  and  $\mathcal{T}'$ , it follows that  $\mathcal{H}_w$  displays  $\mathcal{T}_w$  and  $\mathcal{T}'_w$ . Now  $\mathcal{H}_w$  has  $k$  galls. Suppose that there is a galled-tree network that displays  $\mathcal{T}_w$  and  $\mathcal{T}'_w$ , but has less galls than  $\mathcal{H}_w$ . Then one could use this network to obtain a gall-tree network that displays  $\mathcal{T}$  and  $\mathcal{T}'$  by adjoining the subnetwork below  $w$  in  $\mathcal{H}$  to  $w$  resulting in a galled-tree network with less galls than  $\mathcal{H}$ ; a contradiction to the minimality of  $\mathcal{H}$ . It now follows that amongst all galled-tree networks that display  $\mathcal{T}_w$  and  $\mathcal{T}'_w$ , the galled-tree network  $\mathcal{H}_w$  has the smallest number of galls. By the induction assumption, this implies that  $h(\mathcal{T}_w, \mathcal{T}'_w) = k$  and so

$$\begin{aligned} h(\mathcal{T}, \mathcal{T}') &= h(\mathcal{T}|W, \mathcal{T}'|W) + h(\mathcal{T}_w, \mathcal{T}'_w) \\ &= k + 1. \end{aligned}$$

This completes the proof of the proposition.  $\square$

Nakhleh *et al.* [34] propose a method for inferring hybridization networks that allows for errors in the estimation of the initial two gene trees. In brief, when methods such as maximum likelihood or maximum parsimony infer trees there are a number of equally or close-to-equally good trees that could have been inferred. Thus the strict consensus of each such set of trees is perhaps a better representative of the original data set than one particular tree. However, this representative is typically unresolved, and so an interesting problem is the following. Given two rooted phylogenetic  $X$ -trees  $\mathcal{T}_1$  and  $\mathcal{T}_2$ , determine if there is two rooted binary phylogenetic  $X$ -trees  $\mathcal{T}'_1$  and  $\mathcal{T}'_2$  such that  $\mathcal{T}'_i$  is a refinement of  $\mathcal{T}_i$  with the property that there is a gall-tree network that displays  $\mathcal{T}'_1$  and  $\mathcal{T}'_2$ . Moreover, if there is such a network, find  $\mathcal{T}'_1$  and  $\mathcal{T}'_2$  that minimizes the number of galls over all gall-tree networks that display  $\mathcal{T}'_1$  and  $\mathcal{T}'_2$ . In [34], the authors provide a linear-time algorithm for when the gall-tree network contains exactly one gall. Huynh *et al.* [28] significantly extend this result by providing a quadratic-time algorithm for this problem with no restrictions on the number of galls in the result gall-tree network. Moreover, they also show that this algorithm easily extends to an efficient algorithm for an arbitrary number of input trees. For further details, we refer the reader to [28].

## 5. RECOMBINATION NETWORKS

The *perfect phylogeny with recombination* is a problem that has a very similar flavour to that of MINIMUM HYBRIDIZATION. Indeed, the two problems are closely related. Instead of inputting a collection of trees, the input for this problem is a collection,  $B$  say, of binary sequences. However, the goal is essentially the same. Loosely speaking, this goal is to compute the minimum number of “recombination” events to “explain”  $B$ . Introduced by Hein [21, 22], there is now a number of papers on this problem, including [5, 15, 16, 17, 18, 41, 43, 44, 45, 46]. In this section, we describe this problem and its relationship with MINIMUM HYBRIDIZATION. This relationship will be used in Section 7.

An  $(n, m)$ -*recombination network*  $\mathcal{N}$  is a rooted acyclic digraph with exactly  $n$  vertices of outdegree zero in which each vertex other than the root has either one or two incoming arcs, and each vertex of  $\mathcal{N}$  is labelled with a binary sequence of length  $m$ . The sequence labelling the root is called the *root* or *ancestral* sequence. A vertex with two incoming arcs is called a *recombination* vertex. Each integer in  $\{1, 2, \dots, m\}$  is assigned to exactly one arc of  $\mathcal{N}$  that is not directed towards a recombination vertex. Beginning with the root and its associated sequence, each of the binary sequences labelling the other vertices is based on the binary sequence of its parent and the incoming arc (in the case it is a non-recombination vertex) or its parents (in the case it is a recombination vertex). In particular, the sequences satisfy the following properties:

- (i) If  $v$  is a non-recombination vertex with incoming arc  $e$ , then the sequence labelling  $v$  is obtained from the sequence labelling its parent by changing the  $i$ -th element (site) from 0 to 1 or 1 to 0 appropriately for each integer  $i$  assigned to  $e$ . If no integer is assigned to  $e$ , then the sequence labelling  $v$  is the same as its parent.
- (ii) If  $v$  is a recombination vertex, then, for some positive integer  $p$  strictly between 1 and  $m$  (that is,  $1 < p < m$ ), the sequence labelling  $v$  is the concatenation of the first  $p$  elements of the sequence labelling one of its parents and the last  $m - p$  elements of its other parent. To describe the corresponding recombination event one labels the incoming arcs either  $P$  or  $S$  depending upon which parent contributes the prefix part or the suffix part of the sequence, respectively, and also labels the recombination vertex with an ordered pair indicating the “break-point”.

Biologically speaking, the mutations in (i) are called *point mutations* and, as each site in the sequence mutates exactly once, we are under the infinite sites model of mutations. The recombination process in (ii) is called a *single-crossover recombination* as there is exactly one break-point in the resulting sequence. Even though this model of recombination is very simple, it is the basis of most applications of coalescent theory to recombining sequences [24].

As an example, a recombination network is shown in Fig. 10, where the root sequence is the all-0 sequence. For each recombination vertex in this example, the first two elements in the associated sequence come from its ‘left’ parent and the

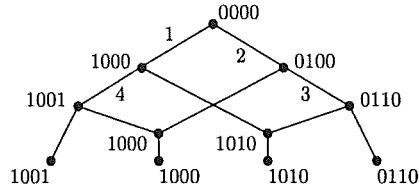


FIGURE 10. A recombination network.

second two elements come from its ‘right’ parent. (We have omitted the labelling of the recombination vertices and their incoming arcs as described in (ii) above.) In the literature, a recombination network is commonly referred to as a “phylogenetic network”.

Let  $B$  be a collection of  $n$  binary sequences of length  $m$ . An  $(n, m)$ -recombination network  $\mathcal{N}$  *explains*  $B$  if the  $n$  vertices of outdegree zero are bijectively labelled with the elements of  $B$ . For example, the recombination network in Fig. 10 explains the collection  $\{1001, 1000, 1010, 0110\}$  of binary sequences. Over all recombination networks that explain  $B$ , we are interested in finding one that has the minimum number of recombination vertices. The perfect phylogeny with recombination problem is formally stated as follows.

#### PERFECT PHYLOGENY WITH RECOMBINATION

**Instance:** A set  $B$  of  $n$  binary sequences of length  $m$ .

**Goal:** Find an  $(n, m)$ -recombination network  $\mathcal{N}$  that explains  $B$  with minimum number of recombination vertices.

**Measure:** The number of recombination vertices in  $\mathcal{N}$ .

Depending upon whether the root sequence of the recombination network is specified or not specified in advance, the problem can be interpreted in one of two ways. In the case that the root sequence is specified in advance, no generality is lost in always choosing the root sequence to be the all-0 sequence. We denote the minimum values for the two problems by  $r(B)$  and  $r^*(B)$ , respectively, and note that  $r^*(B) \leq r(B)$ .

Recombination events are one of the primary influences on genetic variation amongst individuals of the same population. Recognizing how many and where in the sequence these events occur is expected to be a contributing factor in answering a number of important problems in genetics including those centred around genetic diseases. Thus the motivation for PERFECT PHYLOGENY WITH RECOMBINATION is similar to that for MINIMUM HYBRIDIZATION except that our input is now a collection of binary sequences. SNP (single nucleotide polymorphism) sequences satisfy this criteria and are now of great interest (for example, see [25]). Each sequence represents an individual of the same population and, in such a sequence, each site represents an allele of the species. In the case the root sequence is specified in advance, a 0 denotes the ancestral allele, while a 1 denotes the derived (mutant) allele. Observe that  $0 \rightarrow 1$  is the only allowable transition in this case. The reason for the wording “perfect phylogeny” is that the classical perfect phylogeny problem

can be interpreted as the problem of deciding if there is a recombination network with no recombination vertices that explains  $B$ . Without going into any detail, a variation of the above problem that allows for “multiple-crossover” recombinations (more than one break-point) in the resulting recombination networks has also been considered (see [15, 16]).

There is a close relationship between MINIMUM HYBRIDIZATION and PERFECT PHYLOGENY WITH RECOMBINATION with the root sequence specified in advance. In particular, the former problem can be interpreted as a particular instance of the latter.

Using the construction in [46], let  $\mathcal{T}$  and  $\mathcal{T}'$  be two rooted binary phylogenetic  $X$ -trees and let  $|X| = n$ . Noting that  $|E(\mathcal{T})| = |E(\mathcal{T}')| = 2(n-1)$ , bijectively label the edges of  $\mathcal{T}$  and  $\mathcal{T}'$  with the elements of  $\mathcal{C} = \{\chi_1, \chi_2, \dots, \chi_{2(n-1)}\}$  and  $\mathcal{C}' = \{\chi'_1, \chi'_2, \dots, \chi'_{2(n-1)}\}$ , respectively. Each of the elements in  $\mathcal{C}$  and  $\mathcal{C}'$  represent a binary character with states 0 and 1. Associated to each vertex  $v$  (resp.  $v'$ ) of  $\mathcal{T}$  (resp.  $\mathcal{T}'$ ) is the binary sequence of length  $2(n-1)$  in which the  $i$ -th element is 1 if and only if  $\chi_i$  (resp.  $\chi'_i$ ) labels an edge from the root of  $\mathcal{T}$  (resp.  $\mathcal{T}'$ ) to  $v$  (resp.  $v'$ ). Now, for each  $x \in X$ , concatenate the sequences labelling  $x$  in  $\mathcal{T}$  and  $\mathcal{T}'$  with the sequence labelling  $x$  in  $\mathcal{T}'$  following the sequence labelling  $x$  in  $\mathcal{T}$ . Let  $B$  be the resulting collection of  $n$  (concatenated) sequences of length  $4(n-1)$ . The following lemma by Bordewich and Semples [12] provides the above mentioned close relationship.

**Lemma 5.1.** *Let  $\mathcal{T}$  and  $\mathcal{T}'$  be two rooted binary phylogenetic  $X$ -trees, and let  $B$  be the collection of binary sequences that is constructed from  $\mathcal{T}$  and  $\mathcal{T}'$  as above. Then*

$$h(\mathcal{T}, \mathcal{T}') = r(B).$$

The proof of Lemma 5.1 is constructive. In particular, if  $\mathcal{H}$  is a minimum hybridization network that displays  $\mathcal{T}$  and  $\mathcal{T}'$ , then there is a polynomial-time modification of  $\mathcal{H}$  that results in a recombination network  $\mathcal{N}$  that explains  $B$  with the all-0 sequence at the root and has  $h(\mathcal{H})$  recombination vertices. On the other hand, if  $\mathcal{N}$  is a recombination network explaining  $B$  with the all 0-sequence at the root and  $k$  recombination vertices, then  $\mathcal{N}$  can be modified to produce a hybridization network that displays  $\mathcal{T}$  and  $\mathcal{T}'$  with  $k$  hybridization vertices. Again, this modification can be done in polynomial time.

## 6. HYBRIDIZATION NETWORKS IN REAL TIME

An important biological requirement of hybridization networks is that hybridization events occur between contemporaneous taxa (past or present). Maddison [30] pointed out this requirement and, from a mathematical perspective, it has been considered in several papers since including [8, 32, 43, 45]. We begin this section by considering the problem of whether a given hybridization network is consistent with this requirement.



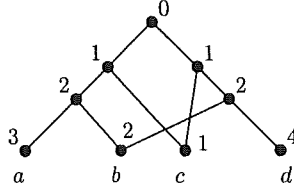


FIGURE 11. A temporal labelling of a hybridization network.

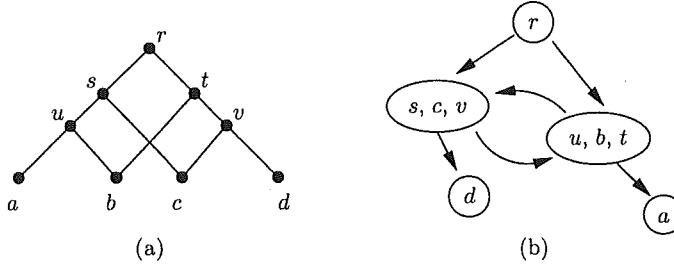


FIGURE 12. (a) A hybridization network with no temporal representation and (b) its temporal digraph.

**6.1. Temporal Representations.** Let  $\mathcal{H}$  be a hybridization network with vertex set  $V$ , and let  $\mathbb{N} = \{0, 1, 2, \dots\}$ . We say that  $\mathcal{H}$  has a *temporal representation* if there is a map  $f : V \rightarrow \mathbb{N}$  that satisfies the following two properties:

- (i) If  $(u, v)$  is an arc of  $\mathcal{H}$  with  $d^-(v) = 1$ , then  $f(u) < f(v)$ .
- (ii) If  $(u, v)$  is an arc of  $\mathcal{H}$  with  $d^-(v) \geq 2$ , then  $f(u) = f(v)$ .

Such a map  $f$  is called a *temporal labelling* of  $\mathcal{H}$ . The purpose of (ii) is so that hybridization events occur with contemporaneous taxa. A temporal labelling of a hybridization network is shown in Fig. 11.

All rooted phylogenetic trees have a temporal representation, but not all hybridization networks have such a representation. For example, the hybridization network in Fig. 12(a) has no temporal representation. The reason for this is that  $u$  and  $t$ , the parents of  $b$ , must coexist in time, while  $s$  and  $v$ , the parents of  $c$ , must also coexist in time. By considering the ancestor-descendant relationships of  $s$  and  $u$ , and  $t$  and  $v$  this is not possible.

We next describe a simple polynomial-time algorithm for deciding whether a hybridization network has a temporal representation and, if so, constructs such a representation. Due to Baroni *et al.* [8], we begin by defining a particular digraph around which the algorithm is based. Let  $\mathcal{H}$  be a hybridization network with vertex set  $V$ . Ignoring the direction of the arcs of  $\mathcal{H}$ , set

$$[v] = \{v\} \cup \{u \in V : \text{there is a path of hybridization arcs from } u \text{ to } v\}.$$

Note that we have partitioned  $V$  into equivalence classes, where  $[v] = \{v\}$  precisely if  $v$  is not incident with a hybridization arc. Setting  $[V] = \{[v] : v \in V\}$ , we define the *temporal digraph* of  $\mathcal{H}$  as the digraph whose vertex set is  $[V]$  and where  $[u]$  and  $[v]$  are joined by an arc  $([u], [v])$  if there is a vertex  $a$  in  $[u]$  and a vertex  $b$  in  $[v]$  such that  $(a, b)$  is an arc of  $\mathcal{H}$  with  $d^-(b) = 1$ . For example, the digraph in Fig. 12(b) is the temporal digraph of the hybridization network in Fig. 12(a).

It turns out that  $\mathcal{H}$  has a temporal representation if and only if its temporal digraph is acyclic and this is the basis of the following algorithm whose correctness is shown in [8].

**Algorithm:** TEMPREP ( $\mathcal{H}$ )

**Input:** A hybridization network  $\mathcal{H}$  with vertex set  $V$ .

**Output:** A temporal labelling of  $\mathcal{H}$  or the statement  $\mathcal{H}$  has no temporal labelling.

1. Construct the temporal digraph  $D_{\mathcal{H}}$  of  $\mathcal{H}$ .
2. Find an acyclic ordering,  $V_0, V_1, \dots, V_k$  say, of  $D_{\mathcal{H}}$ . If there is no such ordering, then return  $\mathcal{H}$  has no temporal representation.
3. Define  $f : V \rightarrow \mathbb{N}$  by setting  $f(v) = i$  for all  $v \in V_i$ , where  $[v] \in V_i$ .
4. Return the map  $f$ .

If a map  $f$  is returned by the algorithm, then  $f$  is a temporal labelling of  $\mathcal{H}$ . It is important to note that a temporal labelling of a hybridization network is no more than an ordering of when past or present taxa appeared. Consequently, it is the ordering on the vertices of  $V$  that is important and not the actual values.

If one is interested in obtaining, up to isomorphism, all temporal labellings of  $\mathcal{H}$ , then the above algorithm can be easily modified to output a list of all such labellings, where a new labelling is outputted in polynomial time and where two labellings are isomorphic if the relative orderings of the vertices are not the same. Essentially, one selects non-empty subsets of vertices that have indegree zero instead of a single vertex in the process of finding an acyclic ordering. All such orderings result in a distinct temporal labelling and all such labellings can be obtained this way. For further details, see [8].

We end this subsection with the following remark. If a hybridization network  $\mathcal{H}$  does not have a temporal representation, then Moret *et al.* [32] observed that, by allowing for missing taxa, one could resolve this issue without adding to the hybridization number of  $\mathcal{H}$ . For example, consider the hybridization network in Fig. 12(a). By creating two new vertices that subdivide the arcs  $(t, b)$  and  $(s, c)$ , and joining pendant arcs to each of these new vertices with new taxa, the resulting hybridization network has a temporal representation. The role of such taxa is to carry a gene or combination of genes from the past into some time when it can be passed on into the new hybrid species. Of course, whether such taxa exist or existed is a separate question.

**6.2. Time Ordered Rooted Subtree Prune and Regraft Operations.** Realizing the importance that time places on possible scenarios for evolutionary histories, Song and Hein [43, 45] (also see [24]) considered a more restrictive notion of the rooted subtree prune and regraft operation. This restriction allows one to attack the problem of PERFECT PHYLOGENY WITH RECOMBINATION in which the root sequence is not specified in advance using rooted subtree prune and regraft operations.

Let  $\mathcal{T}$  be a rooted binary phylogenetic tree and let  $\hat{V} = \{v_1, v_2, \dots, v_{n-2}\}$  be the set of interior vertices of  $\mathcal{T}$ . A *total ordering* on  $\hat{V}$  is a binary relation  $<_{\mathcal{T}}$  given by  $v_i <_{\mathcal{T}} v_j$  if the hypothetical ancestor or speciation event represented by  $v_i$  predates the hypothetical ancestor or speciation event represented by  $v_j$ . In mathematics, total orderings are also called *linear orderings*. We say that  $\mathcal{T}$  is *ordered* if  $\hat{V}$  is totally ordered. By default, such an ordering must preserve the ancestor-descendant relationships given by the topology of  $\mathcal{T}$ .

In performing a rooted subtree prune and regraft operation on an ordered tree  $\mathcal{T}$  one must preserve the ordering on  $\hat{V}$ . In particular, referring to the notation in the definition of a rSPR operation in Section 3, for all  $v_i, v_j \in \hat{V} - \{u\}$ , we have that  $v_i <_{\mathcal{T}'} v_j$  precisely if  $v_i <_{\mathcal{T}} v_j$ , where  $u$  is the “parent” vertex of the root of the subtree being pruned,  $\mathcal{T}$  is the initial tree, and  $\mathcal{T}'$  is the tree resulting from the rSPR operation. Given two ordered rooted binary phylogenetic  $X$ -trees, there is a sequence of (ordered) rSPR operations that transforms one tree into the other. For further combinatorial results on this operation and the ordinary rSPR operation, see Song [39, 40].

Now let  $B$  be a collection of binary sequences of equal length  $m$ . For each  $i$ , the  $i$ -th sites in the sequences induce a character  $\chi_i$ . Under the infinite-sites model of mutation, let  $\mathcal{P}_i$  be the collection of ordered rooted binary phylogenetic  $X$ -trees that display  $\chi_i$ , that is  $\mathcal{P}_i$  is the collection of all such trees for which there exists an edge whose deletion induces the bipartition of  $X$  induced by the character states  $\chi_i$ . Consider the problem of minimizing the following sum:

$$(2) \quad \sum_{i=1}^{m-1} d_{\text{rSPR}}(\mathcal{T}_i, \mathcal{T}_{i+1}),$$

where  $\mathcal{T}_i \in \mathcal{P}_i$  for all  $i$  and  $d_{\text{rSPR}}(\mathcal{T}_i, \mathcal{T}_{i+1})$  denotes the minimum number of (ordered) rSPR operations to transform  $\mathcal{T}_i$  into  $\mathcal{T}_{i+1}$ . It turns out that the minimum value of this sum is equal to  $r^*(B)$ , the optimal value of PERFECT PHYLOGENY WITH RECOMBINATION in which the root sequence is not specified in advance [42]. Thus  $r^*(B)$  can be written in terms of the rSPR distance on ordered rooted binary phylogenetic trees. Moreover, a lower bound for  $r^*(B)$  can be obtained by interpreting the terms in the sum in (2) as the ordinary rSPR distance between two rooted binary phylogenetic trees, where the total ordering on the interior vertices is ignored.

The number of ordered rooted binary phylogenetic trees grows significantly faster than the number of (ordinary) rooted binary phylogenetic trees, and so as it currently stands the above approach to computing  $r^*(B)$  exactly is limiting in practice.

Nevertheless, by studying a particular data set for which previous lower bounds have been calculated, Song and Hein have shown it can work. For further details, see [43, 45] and note that Song and Hein use the terminology “ancestral recombination graph” instead of recombination network.

## 7. COMPUTATIONAL COMPLEXITY

In this section, we discuss some of the computational issues associated with the three main problems that we have discussed in this chapter, namely MINIMUM HYBRIDIZATION, MINIMUM rSPR, and PERFECT PHYLOGENY WITH RECOMBINATION. Throughout this section, the interpretation of the last of these problems will always be the one in which the root sequence is specified in advance.

The following theorem, which we have alluded to several times in this chapter, is due to Bordewich and Semple [11, 12].

**Theorem 7.1.** *Each of the optimization problems MINIMUM HYBRIDIZATION, MINIMUM rSPR, and PERFECT PHYLOGENY WITH RECOMBINATION is NP-hard.*

The proofs of the NP-hardness of MINIMUM HYBRIDIZATION and MINIMUM rSPR make use of their characterizations in terms of agreement forests and use ideas originating from Hein *et al.* [23]. The NP-hardness of PERFECT PHYLOGENY WITH RECOMBINATION follows from Lemma 5.1 and the polynomial-time constructions mentioned after it. To avoid repetition, these comments are also valid for Theorem 7.2.

Despite the negativity of Theorem 7.1, there are some positive results for MINIMUM rSPR. Fixed-parameter algorithms are a practical way to find optimal solutions of NP-hard problems if the parameter measuring the hardness of the problem is small. For MINIMUM rSPR, Bordewich and Semple [11] showed that there is such an algorithm where the rSPR distance itself is the parameter. In particular, instead of computing the rSPR distance between two rooted binary phylogenetic  $X$ -trees by an exhaustive search resulting in an algorithm that takes time  $O((2n)^{2k})$  where  $n = |X|$  and  $k = d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}')$ , they showed that there is a parameterized algorithm for computing this distance in  $O(f(k) \cdot p(n))$  where  $f(k)$  is some computable function depending on  $k$  and  $p$  is a polynomial in  $n$ . The important point of this running time is that  $n$  and  $k$  are now separated which means that, provided  $k$  is small, computing  $d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}')$  may be efficiently possible even when  $n$  is large. The important part of the analysis is Theorem 4.1. For further details of this algorithm and an analysis of how well it works in practice, we refer the interested reader to [11] and [29], respectively.

Translating the setting in [20], Hallet and Lagergren give a fixed-parameter algorithm for a problem that is a restriction of MINIMUM HYBRIDIZATION (also see [1]). Whether there is a fixed-parameter algorithm for MINIMUM HYBRIDIZATION in general is an open problem. For those wanting to find out more about fixed-parameter

algorithms, we refer the reader to [14] and [27]. The latter is an easy-to-read introduction to fixed-parameter algorithms and describe three techniques for developing such algorithms.

For computationally hard problems, polynomial-time approximation algorithms can efficiently find feasible solutions that are sometimes arbitrarily close to the optimal solution. In particular, for a minimization problem, an  $r$ -approximation algorithm means that, for all instances, the size of the feasible solution outputted by the algorithm and the size of an optimal solution is at most  $r$ . The existence of polynomial-time approximation algorithms varies greatly amongst NP-hard problems. For example, regardless of the choice of  $r$ , there is no such algorithm for the general travelling salesman problem unless  $P=NP$ , while for some problems  $\pi$ , no matter how close  $r$  is to 1, there is always such an algorithm. In this latter case, we say that  $\pi$  exhibits a *polynomial-time approximation scheme* (PTAS). Theorem 7.2 is due to Bordewich and Semple [12].

**Theorem 7.2.** *Each of the optimization problems MINIMUM HYBRIDIZATION, MINIMUM RSPR, and PERFECT PHYLOGENY WITH RECOMBINATION is APX-hard. In particular, for each of these problems there is no polynomial-time approximation scheme unless  $P=NP$ .*

For each of our optimization problems, the implication of Theorem 7.2 is that, unless  $P=NP$ , there is some fixed constant  $r$  strictly bigger than 1 for which there is no polynomial-time  $r$ -approximation algorithm. It is shown in [12] that, for each of these problems,  $r$  is at least  $\frac{2113}{2112}$ .

Two polynomial-time approximation algorithms for MINIMUM RSPR have appeared in the literature [23, 35]. Both are stated as 3-approximation algorithms, however, each of these algorithms have been subsequently shown to be incorrect in some way. Nevertheless, using these approaches, Bonet *et al.* [10] describe a polynomial-time 5-approximation algorithm for MINIMUM RSPR. Intuitively, this algorithm builds an agreement forest locally. Currently, there appears to be no such algorithm for MINIMUM HYBRIDIZATION. One might hope that the algorithm in [10] extends to MINIMUM HYBRIDIZATION, but, due to the additional global condition on an acyclic-agreement forest, it seems unlikely that such an approach will work. For an excellent reference on approximation algorithms, see [4].

## 8. CONCLUDING COMMENTS

The understanding and analysis of reticulation in evolution is playing a prominent role in modern-day phylogenetics. In this chapter, we considered one particular, but central, aspect; namely the problem of finding the smallest number of reticulation events that are required to explain the evolution of a collection of species under consideration subject to some initial input. For us, the input was a collection of rooted phylogenetic trees. The approach we have taken here is analytical so as to provide a theoretical foundation for algorithmic solutions to the problem. Furthermore, our main interest has been on a general solution rather than

one that is restricted in some way. Unfortunately, despite the fixed-parameter algorithm for MINIMUM RSPR and the divide-and-conquer algorithm for MINIMUM HYBRIDIZATION described in this chapter, we are always going to be limited in finding exact solutions because of the NP-hardness of these problems. This turns our attention to future work.

A number of papers have considered efficient algorithms for computing lower bounds for PERFECT PHYLOGENY WITH RECOMBINATION (for example, see [5, 19, 26, 31, 44]). While one could use the constructions outlined after Lemma 5.1 and these results, it appears that little attention has been given to finding such algorithms directly for MINIMUM RSPR and MINIMUM HYBRIDIZATION. Given the incorrectness of previous approximations for MINIMUM RSPR, a mathematically challenging task is to improve the 5-approximation algorithm for this problem. Whether MINIMUM HYBRIDIZATION even has such an algorithm, regardless of the size of the ratio, is an interesting question. While we have only considered combinatorial questions in this chapter, it is statistical questions that will eventually need to be addressed. For example, how can one use differing bootstrap support values for conflicting phylogenies to quantify and distinguish between genuine reticulation and other biological processes that give rise to conflicts such as lineage sorting? Combinatorial considerations are often the first steps towards statistical-based approaches in phylogenetics and so it is likely that combinatorial insights into hybridization networks will also help in the development of such approaches to reticulation.

#### ACKNOWLEDGMENTS

Many thanks to Katherine St. John and Yun Song for a number of helpful discussions during the writing of this chapter. This work was supported by the New Zealand Marsden Fund.

#### REFERENCES

- [1] Addario-Berry, L., Hallett, M., and Lagergren, J. (2003). Towards identifying lateral gene transfer events. In: *Proceedings of the Pacific Symposium on Biocomputing*, Vol. 8, pp. 279-290.
- [2] Allan, H. H. (1961). *Flora of New Zealand, Volume I, Indigenous tracheophyta: Psilopsida, Lycopsida, Filicopsida, Gymnospermae, Dicotyledones*. Government Printer, Wellington, New Zealand.
- [3] Allen, B. L. and Steel, M. (2001). Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, 5, 1-13.
- [4] Ausiello, G., Crescenzi, P., Gambosi, G., Kann, V., Marchetti-Spaccamela, A., and Protasi, M. (1999). *Complexity and Approximation*. Springer, Berlin.
- [5] Bafna, V. and Bansal, V. (2004). The number of recombination events in a sample history: conflict graph and lower bounds. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1, 78-90.
- [6] Baroni, M. (2004). Hybrid phylogenies: a graph-based approach to represent reticulate evolution. Unpublished PhD thesis, University of Canterbury.
- [7] Baroni, M., Semple, C., and Steel, M. (2004). A framework for representing reticulate evolution. *Annals of Combinatorics*, 8, 391-408.
- [8] Baroni, M., Semple, C., and Steel, M. (2006). Hybrids in real time. *Systematic Biology*, 55, 46-56.

- [9] Baroni, M., Grünewald, S., Moulton, V., and Semple, C. (2005). Bounding the number of hybridization events for a consistent evolutionary history. *Mathematical Biology*, 51, 171-182.
- [10] Bonet, M. K., St. John, K., Mahindru, R., and Amenta, N. (2006). Approximating subtree distances between phylogenies. Technical Report #669, Centre de Recerca Matemàtica, Barcelona.
- [11] Bordewich, M. and Semple, C. (2004). On the computational complexity of the rooted subtree prune and regraft distance. *Annals of Combinatorics*, 8, 409-423.
- [12] Bordewich, M. and Semple, C. Computing the minimum number of hybridisation events for a consistent evolutionary history, submitted.
- [13] Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science*, 284, 2124-2128.
- [14] Downey, R. and Fellows, M. (1998). *Parameterized Complexity*. Springer, New York.
- [15] Gusfield, D. (2005). Optimal, efficient reconstruction of root-unknown phylogenetic networks with constrained and structured recombination. *Journal of Computer and System Sciences*, 70, 381-398.
- [16] Gusfield, D. and Bansal, V. (2005). A fundamental decomposition theory for phylogenetic networks and incompatible characters. In: *Proceedings of the Ninth Annual International Conference on Research in Computational Molecular Biology (RECOMB 2005)* (ed. S. Miyano et al.), Lecture Notes in Bioinformatics, Vol. 3500, Springer, Berlin, pp. 217-232.
- [17] Gusfield, D., Eddhu, S., and Langley, C. (2004). Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *Journal of Bioinformatics and Computational Biology*, 2, 173-213.
- [18] Gusfield, D., Eddhu, S., and Langley, C. (2004). The fine structure of galls in phylogenetic networks. *INFORMS Journal on Computing*, 16, 459-469.
- [19] Gusfield, D., Hickerson, D., and Eddhu, S. An efficiently-computed lower bound on the number of recombinations in phylogenetic networks: theory and empirical study. *Discrete Applied Mathematics*, in press.
- [20] Hallett, M. and Lagergren, J. (2001). Efficient algorithms for lateral gene transfer problems. In: *Proceedings of the Fifth Annual International Conference on Research in Computational Molecular Biology (RECOMB 2001)*, ACM Press, New York, pp. 149-156.
- [21] Hein, J. (1990). Reconstructing evolution of sequences subject to recombination using parsimony. *Mathematical Biosciences*, 98, 185-200.
- [22] Hein, J. (1993). A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution*, 36, 396-405.
- [23] Hein, J., Jing, T., Wang, L., and Zhang, K. (1996). On the complexity of comparing evolutionary trees. *Discrete Applied Mathematics*, 71, 153-169.
- [24] Hein, J., Schierup, M., and Wiuf, C. (2005). *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press.
- [25] Hinds, D., Stuve, L., Nilsen, G., Halperin, E., Eskin, E., Gallinger, D., Frazer, K., and Cox, D. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science*, 307, 1072-1079.
- [26] Hudson, R. and Kaplan, N. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111, 147-164.
- [27] Hüffner, F., Niedermeier, R., and Wernick, S. Techniques for practical fixed-parameter algorithms, submitted.
- [28] Huynh, T. N. D., Jansson, J., Nguyen, N. B., and Sung, W. -K. (2005). Constructing a smallest refining galled phylogenetic network. In: *Proceedings of the Ninth Annual International Conference on Research in Computational Molecular Biology (RECOMB 2005)* (ed. S. Miyano et al.), Lecture Notes in Bioinformatics, Vol. 3500 Springer, Berlin, pp. 265-280.
- [29] Linz, S., St. John, K., and Semple, C. Experimental and theoretical analysis of hybridization, in preparation.
- [30] Maddison, W. (1997). Gene trees in species trees. *Systematic Biology*, 46, 523-536.
- [31] Myers, S. and Griffiths, R. (2003). Bounds on the minimum number of recombination events in a sample history. *Genetics*, 163, 375-394.
- [32] Moret, B. M. E., Nakhleh, L., Warnow, T., Linder, C. R., Tholse, A., Padolina, A., Sun, J., and Timme, R. (2004). Phylogenetic networks: modeling, reconstructibility, and accuracy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1, 1-11.

- [33] Nakhleh, L., Ruths, D., and Wang, L. S. (2005). RIATA-HGT: a fast and accurate heuristic for reconstructing horizontal gene transfer. In: *Proceedings of the Eleventh International Computing and Combinatorics Conference (COCOON 05)* (ed. L. Wang), Lecture Notes in Computer Science, Vol. 3595, Springer, pp. 84-93.
- [34] Nakhleh, L., Warnow, T., Linder, C. R., and St. John, K. (2005). Reconstructing reticulate evolution in species—theory and practice. *Journal of Computational Biology*, **12**, 796-811.
- [35] Rodrigues, E. M., Sagot, M. -F., and Wakabayashi, Y. (2001). Some approximation results for the maximum agreement forest problem. In: *Approximation, Randomization and Combinatorial Optimization: Algorithms and Techniques (APPROX and RANDOM)* (ed. M. Goemans et al.), Lecture Notes in Computer Science, Vol. 2129, Springer, Berlin, pp. 159-169.
- [36] Schmidt, H. A. (2003). Phylogenetic trees from large data sets. Unpublished PhD thesis, Heinrich-Heine Universität.
- [37] Semple, C. and Steel, M. (2003). *Phylogenetics*. Oxford University Press.
- [38] Semple, C. and Steel, M. (2006). Unicyclic networks: compatibility and enumeration. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **3**, 84-91.
- [39] Song, Y. S. (2003). On the combinatorics of rooted binary phylogenetic trees. *Annals of Combinatorics*, **7**, 365-379.
- [40] Song, Y. S. (2006). Properties of subtree-prune-and-regraft operations on totally-ordered phylogenetic trees. *Annals of Combinatorics*, **10**, 147-163.
- [41] Song, Y. S. (2006). A concise necessary and sufficient condition for the existence of a galled-tree. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **3**, 186-191.
- [42] Song, Y. S. (2006). Private communication.
- [43] Song, Y. S. and Hein, J. (2003). Parsimonious reconstruction of sequence evolution and haplotype blocks: finding the minimum number of recombination events. In: *Algorithms in Bioinformatics (WABI 2003)* (ed. G. Benson and R. Page), Lecture Notes in Bioinformatics, Vol. 2812, Springer, Berlin, pp. 287-302.
- [44] Song, Y. S. and Hein, J. (2004). On the minimum number of recombination events in the evolutionary history of DNA sequences. *Journal of Mathematical Biology*, **48**, 160-186.
- [45] Song, Y. S. and Hein, J. (2005). Constructing minimal ancestral recombination graphs. *Journal of Computational Biology*, **12**, 147-169.
- [46] Wang, L., Zhang, K., and Zhang, L. (2001). Perfect phylogenetic networks with recombination. *Journal of Computational Biology*, **8**, 69-78.

BIOMATHEMATICS RESEARCH CENTRE, DEPARTMENT OF MATHEMATICS AND STATISTICS, UNIVERSITY OF CANTERBURY, CHRISTCHURCH, NEW ZEALAND

E-mail address: c.semple@math.canterbury.ac.nz